IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Adaptation robuste de modèles HMM pour la vérification du locuteur dépendante du texte

Johnny Mariéthoz [*] Frédéric Bimbot [**]

IDIAP–RR 00-08

June 2000

published in
JEP2000

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[*] IDIAP - BP 592, CH-1920 Martigny, Switzerland
[**] IRISA - Campus Beaulieu, 35042 Rennes, France

IDIAP Research Report 00-08

Adaptation robuste de modèles HMM pour la vérification du locuteur dépendante du texte

Johnny Mariéthoz Frédéric Bimbot

June 2000

published in
JEP2000

Abstract. When deploying a secure system based on speaker verification, the limited amount of training data is usually critical. Indeed, the enrollment procedure must be fast and user-friendly. An incremental training of HMM speaker models, based on a MAP (Maximum A Posteriori) adaptation technique is used in order to make the enrollment more robust with only one or two utterances of the client password. This paper presents the improvements which can be achieved, in term of verification performance and stability of the decision thresholds. Our results highlight the benefits of MAP adaptation in conjunction with a synchronous alignment approach.

Table des matières

1 Introduction **3**

2 Cadre général **3**
2.1 Modèle probabiliste 3
2.2 Décision et types d'erreurs 4
2.3 Mesure des performances 4

3 Apprentissage incrémental **4**
3.1 Modalités d'apprentissage 4
3.2 Adaptation Bayésienne 5

4 Performances **5**
4.1 Protocole d'évaluation 5
4.2 Résultats 5

5 Dérive des seuils **7**
5.1 Analyse diagnostique 7
5.2 Apprentissage par adaptation 7
5.3 Résultats 7
5.4 Alignement synchrone 8

6 Conclusions **9**

Table des figures

1 batch vs incrémental 6
2 Dérive des seuils en mode incrémental 6
3 Dérive des seuils avec une adaptation Bayesienne du modèle du monde avec $\gamma = 2/3$. . 8
4 Dérive des seuils avec une adaptation Bayesienne du modèle Ω avec $\gamma = 2/3$ et alignement synchrone sur Ω à l'apprentissage et au décodage. 9

1 Introduction

La vérification du locuteur suscite un intérêt croissant de la part des fournisseurs de services téléphoniques, dans la mesure où ces techniques permettent de mieux sécuriser les transactions vocales sur les différents réseaux de télécommunications, en offrant la possibilité de réduire les risques de fraude sans nécessiter l'implantation d'équipement supplémentaire chez l'abonné. Cependant, des difficultés spécifiques existent pour ce type d'applications commerciales, une d'entre elle étant la nécessité de garantir une mise en oeuvre rapide du service pour tout nouvel utilisateur. En pratique, les applications visées doivent être opérationnelles à partir d'une ou deux sessions d'entraînement, ce qui limite considérablement la représentativité des données d'apprentissage, que ce soit en termes de couverture de la variabilité individuelle au cours du temps ou de type de microphone et de canal de transmission observé.

Pour remédier à ce problème, une solution consiste à affiner, au fil de l'utilisation du système, les modèles caractéristiques de chaque client avec les énoncés produits par ce client à l'occasion d'utilisation précédentes du service, afin d'acquérir progressivement des données plus représentatives des différentes conditions d'utilisation de l'application par ce client.

Les travaux présentés dans cet article se placent dans le contexte d'un formalisme probabiliste du problème de la vérification du locuteur, où la décision est prise à partir d'un rapport de vraisemblance fourni par le modèle spécifique du client et un modèle indépendant du locuteur (appelé *modèle du monde*).

Nous utilisons les techniques d'adaptation Bayésienne pour effectuer l'apprentissage incrémental du modèle du client. Nous comparons tout d'abord l'impact, sur les performances du système, d'une approche incrémentale par adaptation à partir des données nouvelles par rapport à une approche par réapprentissage complet utilisant l'ensemble des données produites. Nous commentons ensuite nos observations sur la dérive des seuils de décision optimaux et nous présentons une solution permettant de remédier aux problèmes rencontrés, en utilisant également une technique Bayésienne pour estimer le modèle client initial. Enfin, nous mettons en évidence un avantage supplémentaire à utiliser une technique d'alignement synchrone pour calculer le rapport de vraisemblance sur une séquence d'états commune aux modèles du client et du monde.

Les travaux rapportés dans cet article sont effectués dans le contexte du projet Européen Telematics PICASSO [B$^+$99] (Work-Package 5). Les expériences ont été réalisées avec la plate-forme logicielle commune *Picassoft* sur la base de données PolyVar / suisse romand selon un protocole expérimental défini par l'ensemble des partenaires.

2 Cadre général

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 enoncé 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 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. Quand le modèle du non-client est le même pour tous les clients, ce qui est le cas dans ces travaux, on le désigne par *modèle du monde* (noté Ω).

Dans les travaux décrits ici, la vérification s'effectue sur un mot (ou un groupe de mot) issu d'un vocabulaire de 17 mots différents, ce vocabulaire étant commun à tous les clients. Les modèles probabilistes utilisés sont des HMM (Modèles de Markov Cachés) à topologie gauche-droite (un par mot) dont les fonctions d'émission des états sont des mélanges de distributions Gaussiennes. Une spécificité importante de nos HMM réside en ce que les modèles client et le modèle du monde ont une topologie identique.

2.2 Décision et types d'erreurs

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.

2.3 Mesure des performances

Les performances des approches décrites dans cet article sont présentées sous deux formes:

- une courbe DET [MP97] qui indique 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 est le système.
- les performances du système dans une condition de fonctionnement équicoût, c'est-à-dire pour laquelle les deux types d'erreur sont considérées comme étant de gravité égale. Dans ce cas, la décision optimale vise à minimiser le Demi Taux d'Erreur Total (DTET), c'est-à-dire la moyenne arithmétique du taux de faux rejets et du taux de fausses acceptations. Les résultats présentés sont obtenus par réglage des seuils a posteriori c'est-à-dire en optimisant le DTET sur l'ensemble de test. Notons que, pour la condition de fonctionnement équicoût, le seuil Bayésien théorique sur le *logarithme* du rapport de vraisemblance est égal à 0.

3 Apprentissage incrémental

3.1 Modalités d'apprentissage

Une partie de notre étude consiste à comparer les performances du système selon deux modalités d'apprentissage des modèles du client, désignées par mode *batch* et mode *incrémental*.

Dans les deux modalités, un modèle client initial est estimé à partir de 1 répétition, provenant des 2ères sessions d'enregistrement, soit 2 énoncés au total. Le modèle initial ainsi obtenu sera désigné dans la suite par l'abréviation '12'. Précisons que l'algorithme d'apprentissage utilisé est l'algorithme des k-moyennes segmentales, c'est-à-dire un algorithme EM avec segmentation par Viterbi, où le modèle initial est le modèle du monde.

Dans le mode *batch*, on réestime complètement, après chaque nouvelle session, le modèle client à partir des données d'initialisation (2ères sessions) auxquelles on adjoint successivement la répétition des sessions ultérieures (session 3, puis 4, puis 5), soit des ensembles d'apprentissage constitués respectivement de 3, 4 et 5 énoncés. Chaque réestimation nécessite que soit conservés en mémoire les énoncés représentés sous forme acoustique (paramétrée). On désignera ces configurations par les abréviations 123, 1234 et 12345.

Dans le mode *incrémental*, on fait l'hypothèse que l'on a plus accès aux données acoustiques des sessions passées et que l'on doit se limiter à adapter le modèle client courant à partir du seul énoncé de la session précédente. Cette contrainte est imposée par la préoccupation de minimiser et de contrôler le volume occupé par les informations nécessaires à caractériser le client. On désignera par 12+3, 12+3+4 et 12+3+4+5 les configurations correspondant à l'apprentissage incrémental avec les sessions 3, 4 et 5 respectivement.

3.2 Adaptation Bayésienne

Les techniques d'adaptation Bayésienne sont couramment utilisées pour l'estimation statistique des modèles probabilistes utilisés en reconnaissance de la parole [GL94] car elles offrent un cadre théorique et une bonne efficacité pratique pour traiter des différents problèmes d'adaptation que l'on peut rencontrer dans les contextes applicatifs, que ce soit l'adaptation au locuteur, au canal, à l'environnement d'utilisation, etc (voir par exemple [MC99]).

Nous adoptons cette même approche pour l'apprentissage incrémental: nous considérons que le problème posé revient à adapter le modèle client estimé sur les sessions initiales, à partir de nouvelles observations (en l'occurrence les données client provenant des sessions ultérieures) [Mok98]. Par ailleurs, on se place dans le cadre de l'adaptation supervisée, c'est-à-dire sous l'hypothèse que les données servant à adapter le modèle proviennent effectivement du client. Des travaux parallèles [F^+00] étudient le comportement de l'apprentissage incrémental en cas d'attaques d'imposteurs.

En pratique, nous utilisons une version simplifiée de l'apprentissage Bayésien qui consiste à n'actualiser que les moyennes des distributions gaussiennes, selon la formule d'adaptation:

$$\mu_{n+1} = \frac{\alpha_n \, \mu_n \; + \; a\, m}{\alpha_n \; + \; a}$$

où μ_n et μ_{n+1} désignent respectivement les moyennes du modèle avant et après adaptation et où m représente la moyenne des données observées. Le poids α_n est pris égal au nombre de données utilisées pour estimer la valeur de μ_n et a correspond au nombre de valeurs observées pour calculer m. A chaque incrément, α est remis à jour: $\alpha_{n+1} = \alpha_n + a$.

4 Performances

4.1 Protocole d'évaluation

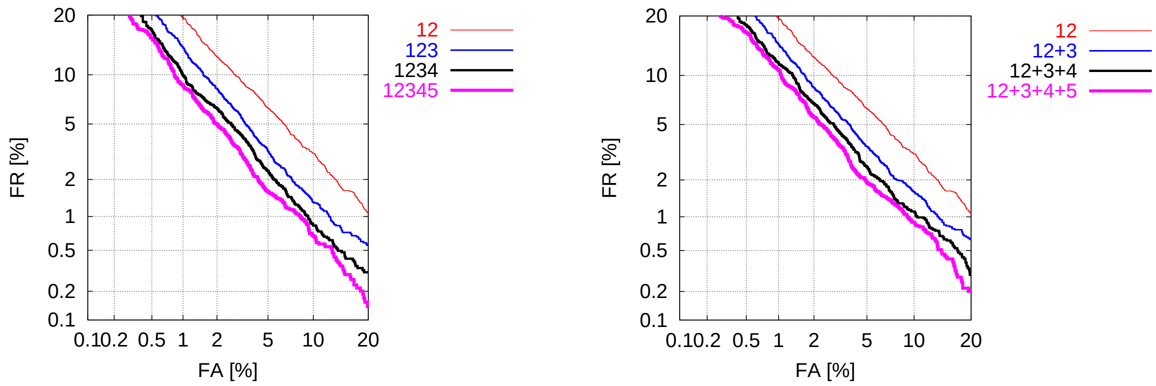
La base de données utilisée comporte 17 mots de commande[1] provenant de la base de données PolyVar / suisse romand. La population des clients est constituée de 19 locuteurs (12 hommes et 7 femmes). Une autre population de 56 locuteurs (28 hommes et 28 femmes) est utilisées pour estimer le modèle du monde (56 énoncés). Les résultats expérimentaux sont obtenus à partir d'environ 6000 accès clients (soit, en moyenne, de l'ordre de 15 accès par client et par mot) et d'à peu près 12000 accès imposteurs (issus de la même population que celle des clients).

Les coefficients LPCC d'ordre 16, ainsi que les deltas et les delta-deltas sont utilisés pour la paramétrisation acoustique des énoncés. La topologie des modèles HMM des clients et du monde est identique, à savoir 2 états par phonème et 1 gaussienne par état.

4.2 Résultats

La figure 1 présente, sous forme de courbes DET, les performances des deux protocoles d'apprentissage (incrémental vs batch). Ces figures mettent en évidence un avantage relativement marginal de l'approche batch. On retiendra donc que l'approche incrémentale ne semble pas dégrader les performances de façon sensible et qu'il est donc judicieux de l'utiliser dès lors que les capacités de stockage pour chaque client sont limitées.

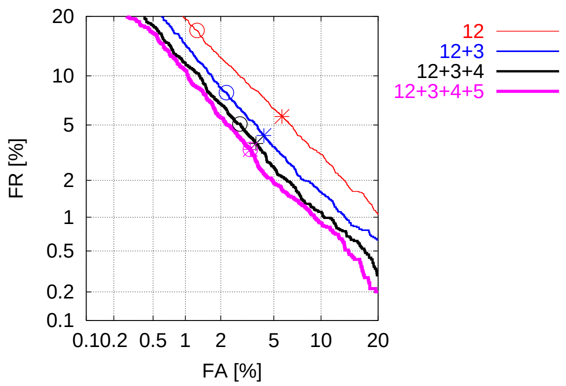
1. annulation, casino, cinéma, concert, corso, exposition, galerie du Manoir, Gianadda, guide, Louis Moret, Manifestation, message, mode d'emploi, musée, précédent, quitter, suivant



(a) batch (b) incrémental

Sessions	12	12+3	12+3+4	12+3+4+5
DTET [%] batch	5.67	4.07	3.61	3.12
DTET [%] incremental	5.67	4.26	3.73	3.39

FIG. 1 – *batch vs incrémental*



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	-0.678	-0.353	-0.201	-0.092

FIG. 2 – *Dérive des seuils en mode incrémental*

En revanche, on observe, pour les deux méthodes, une dérive du seuil optimal pour le DTET en fonction du nombre de sessions prises en compte. Ceci est mis en évidence, pour l'apprentissage incrémental, sur la figure 2 où l'on peut comparer, les points de fonctionnements optimaux pour chaque configuration (représentés pas des croix) et le point de fonctionnement correspondant à un seuil fixe, optimisé sur la configuration à 5 sessions (représentés par un rond). Le même phénomène s'observe dans le cas de l'apprentissage batch. Cette dérive est génante, car elle rend nécessaire une estimation de seuil différente pour chaque réestimation ou chaque adaptation.

5 Dérive des seuils

5.1 Analyse diagnostique

Une analyse fine du comportement du système sur la tâche traitée indique que la dérive des seuils provient de la variation dans la qualité de l'estimation des modèles du client, estimation d'autant plus mauvaise que le volume de données utilisé est limité. En effet, en cas de données insuffisantes, les estimateurs $\hat{\mathcal{P}}(Y|X)$ des distributions clients $\mathcal{P}(Y|X)$ sont de très mauvaise qualité et induisent de ce fait un biais négatif dans la valeur du seuil optimal, par rapport au seuil Bayésien théorique (égal à 0 pour le DTET).

D'un point de vue pratique, ces mauvaises estimations se manifestent à deux niveaux: d'une part une mauvaise estimation des moyennes des gaussiennes dans les états du HMM client. D'autre part, un chemin de décodage inadéquat lors de l'alignement de l'énoncé de test avec le modèle client. Les scores de vraisemblance de chaque trame de test sont donc doublement entâchés d'erreur.

5.2 Apprentissage par adaptation

L'approche utilisée dans les expériences précédentes repose sur un apprentissage du modèle client initial (configuration 12) à partir des données d'entraînement correspondantes, en utilisant le modèle Ω comme initialisation de l'algorithme EM. Néanmoins, au cours des itérations, certains états peuvent devenir faiblement occupés, voire totalement désertés par les données d'apprentissage, ce qui a une influence néfaste tant sur les capacités de généralisation du modèle que sur la qualité de l'alignement qu'il peut fournir sur de nouvelles observations.

C'est pourquoi nous avons opté dans notre contexte de vérification *dépendante* du texte pour une approche d'estimation des modèles clients initiaux basée sur l'adaptation Bayésienne du modèle du monde, s'appuyant sur des résultats montrant l'intérêt de procéder ainsi en vérification du locuteur *indépendante* du texte [Rey97]. Seule l'intialisation du modèle client est modifée (étape 12), ensuite l'apprentissage reste identique (étape 12+3, etc).

En utilisant le même formalisme d'adaptation que précédemment, l'approche proposée revient donc à estimer les moyennes des gaussiennes des fonctions d'émission pour les HMM clients sous la forme:

$$\mu_X = \frac{\beta\,\mu_\Omega \;+\; b\,m_X}{\beta \;+\; b}$$

où μ_X est la moyenne de la gaussienne du modèle client adapté, μ_Ω la moyenne de la gaussienne correspondante dans le modèle du monde, m_X la moyenne des données client associées à la gaussienne, et β et b les poids attribués au modèle du monde et aux données client respectivement.

Contrairement au cas précédent, les poids β et b ne peuvent être choisis égaux au nombre d'observations associées à la gaussienne considérée dans le modèle du monde et le modèle client respectivement, car, en pratique, $\beta \gg b$. C'est pourquoi on réécrit l'équation précédente sous la forme:

$$\mu_X = \gamma\,\mu_\Omega \;+\; (1-\gamma)m_X$$

et l'on choisit une valeur de γ commune à toutes les gaussiennes de façon à optimiser les performances du système. Le paramètre γ correspond alors au poids relatif apparent du modèle du monde dans le processus d'adaptation. Dans nos expériences, nous avons testé les valeurs de γ égales à 0, 1/2, 2/3 et 3/4.

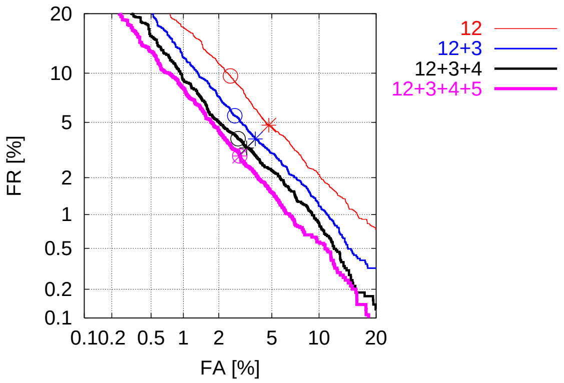
5.3 Résultats

La table 1 montre l'influence des valeurs de γ sur le DTET et sur la valeur du seuil optimal pour le modèle intial et ceux issus d'un apprentissage incrémental.

On observe que c'est pour les valeurs de γ de 1/2 et de 2/3 que les performances optimales sont obtenues avec une dérive des seuils moindre dans le second cas. Notons que dans ce dernier cas, le seuil optimal ne diffère du seuil théorique que de 5 à 15 % selon le nombre de sessions d'adaptation.

TAB. 1 – *Influence de γ sur le modèle initial et sur les modèles obtenus par apprentissage incrémental*

		γ			
		0	1/2	2/3	3/4
12	seuil	−0.818	−0.252	0.053	0.190
	DTET[%]	6.26	4.76	4.85	5.22
12+3	seuil	−0.549	−0.146	0.103	0.239
	DTET[%]	4.15	3.69	3.88	4.15
12+3+4	seuil	−0.371	−0.068	0.059	0.276
	DTET[%]	3.40	3.24	2.88	3.49
12+3+4+5	seuil	−0.238	0.001	0.177	0.290
	DTET[%]	2.96	2.88	2.87	3.16



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	0.072	0.131	0.170	0.208

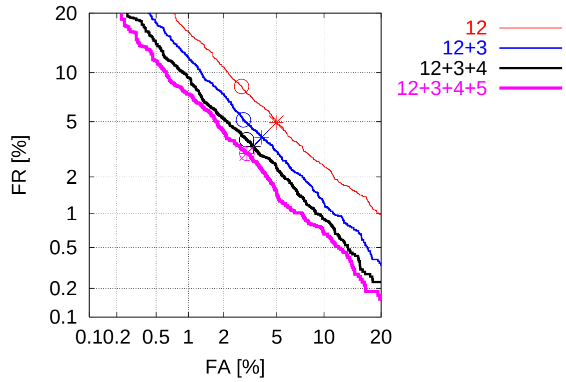
FIG. 3 – *Dérive des seuils avec une adaptation Bayesienne du modèle du monde avec $\gamma = 2/3$.*

5.4 Alignement synchrone

Pour tenter d'accroître la robustesse du système, nous avons intégré dans le processus de vérification une technique de synchronisation des alignements des observations acoustiques dans le modèle client et dans le modèle du monde. Selon cette approche [M^+99], la séquence d'états dans les deux modèles est exactement la même et est, en l'occurence, définie par l'alignement dans le modèle du monde.

La figure 4 montre les résultats avec adaptation dans le cas de l'utilisation d'un alignement synchrone (sur le modèle Ω) pour l'apprentissage et le décodage.

Il est intéressant de noter que les performances en terme de courbes DET sont similaires à celles observées précédemment (figure 3), mais que la dérive du seuil est considérablement réduite, avec une fluctuation du seuil optimal de l'ordre de 5 % seulement autour du seuil théorique. En outre, le temps nécessaire à une vérification est quasiment divisé par deux, car il suffit d'effectuer un seul décodage Viterbi au lieu de deux. Ces résultats supplémentaires confirment donc tout l'intérêt de la technique d'alignement synchrone pour la vérification du locuteur.



Sessions	12	12+3	12+3+4	12+3+4+5
Seuils	-0.044	0	0.035	0.069

FIG. 4 – *Dérive des seuils avec une adaptation Bayesienne du modèle Ω avec $\gamma = 2/3$ et alignement synchrone sur Ω à l'apprentissage et au décodage.*

6 Conclusions

Nos travaux tendent à mettre en évidence l'apport des techniques d'adaptation à différents niveaux de l'apprentissage des modèles de locuteur. Nos expériences illustrent l'intérêt d'adapter le modèle client à partir d'un modèle indépendant du locuteur. Elles valident également l'utilisation d'un apprentissage incrémental permettant de remettre à jour de façon incrémentale le modèle du client à partir des énoncés prononcés en phase opérationnelle, sans avoir à stocker l'ensemble des données acoustiques correspondantes. Enfin, nous confirmons l'intérêt de l'alignement synchrone qui semble contribuer à faciliter le réglage et le suivi des seuils en apprentissage incrémental.

Une des étapes suivantes consiste à étendre cette étude au cas de l'apprentissage non-supervisé, c'est-à-dire sans savoir a priori si les énoncés d'apprentissage ont effectivement été produits ou non par le client.

Remerciements

Ce travail est financé par l'OFES (Office Fédéral de l'Education et de la Science), project n 97.0494-2 et par la CE (Commission Européenne) Telematics Programme LE4 (project 8369).

Références

[B+99] F. Bimbot et al. 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 5–10 1999.

[F+00] C. Fredouille et al. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In *ICASSP2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 5–9 2000.

[GL94] J. L. Gauvain and C.-H. Lee. Maximum a posterionri estimation for multivariate gaussian mixture observation of markov chains. In *IEEE Transactions on Speech Audio Processing*, volume 2, pages 291–298, April 1994.

[M+99] J. Mariéthoz et al. Client / world model synchronous alignement for speaker verification. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, Budapest, Hungary, September 5–10 1999.

[MC99] C. Mokbel and O. Collin. Incremental enrollment of speech recognizers. In *ICASSP'99*, 1999.

[Mok98] C. Mokbel. Incremental enrollment. PICASSO WP5 Delivrable D5.1, December 1998.

[MP97] A. Martin and M. Przybocki. The det curve in assessment of detection task performance. In *Eurospeech 97*, volume 4, pages 1895–1898, 1997.

[Rey97] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech 97*, volume 2, pages 963–966, 1997.