Gémellité et reconnaissance automatique du locuteur

Nicolas Scheffer †, Jean-Francois Bonastre †, Alain Ghio ‡, Bernard Teston ‡

† Laboratoire d'Informatique d'Avignon - Université d'Avignon et des Pays du Vaucluse
339, chemin des Meinajaries - Agroparc BP 1228 84911 AVIGNON Cedex 9 - FRANCE
Mél : {nicolas.scheffer, jean-francois.bonastre}@lia.univ-avignon.fr - http ://www.lia.univ-avignon.fr

‡ Laboratoire Parole et Langage - UMR 6057 CNRS - Université de Provence
29, avenue Robert Schuman 13621 AIX EN PROVENCE Cedex 1 - FRANCE
Mél : {alain.ghio, bernard.teston}@lpl.univ-aix.fr - http ://www.lpl.univ-aix.fr

ABSTRACT

This paper presents a speaker recognition system applied to a limited corpus of twin speakers. Two experiments are carried out using the LIA speaker recognition platform AMIRAL. The aim of the first experiment is to identify a twin of a speaker among a corpus, whereas the second one is a classical speaker verification task. The results show that an automatic system is generally able to identify a twin with an acceptable performance (85 % of good identification). Furthermore, the system is still able to discriminate the target speaker from its twin. For the verification experiment, we obtain 6% F.A. for 0% F.R. and 53% F.R. for 0% F.A.. As the available amount of data is extremely limited, results have to be interpreted with caution.

1. INTRODUCTION

La production vocale de jumeaux n'est pas un sujet d'étude récent [6]. Ce thème reste très délicat car il est difficile de faire la part des choses entre des similarités dues à la gémellité (caractéristiques physiologiques proches héritées d'un patrimoine génétique commun), à la fratrie (habitudes et mimétismes langagiers), à l'environnement sociolinguistique (lieux de vie et cadre de vie souvent communs)... L'aspect psychologique, et particulièrement le rapport de chaque individu par rapport à sa gémellité (rejet ou relation fusionnelle), apparaît comme un facteur essentiel mais difficile à maîtriser. Enfin, ce type d'étude reste difficile du fait de la difficulté de recrutement de cette classe de locuteurs. Ce domaine a cependant été abordé par [7]. Il intéresse non seulement les spécialistes du développement [8], mais aussi les phonéticiens [9] et bien évidemment les professionnels de la reconnaissance automatique du locuteur.
Dans cet article, nous étudions la sensibilité d'un système de Reconnaissance Automatique du Locuteur (RAL) à la présence de jumeaux dans un corpus. Nous disposons pour l'expérimentation de la plate-forme AMIRAL [4] (développée au Laboratoire d'Informatique d'Avignon et basée sur des méthodes statistiques) et d'un corpus de petite taille composé de vrais jumeaux.
La *Section 2* présente brièvement le corpus ainsi que les outils statistiques utilisés (Modèles de Mélanges de Gaussiennes, rapport de vraisemblance). Les expérimentations sont présentées en *Section 3* et *4*. Une première expérience (cf. *Section 3*) consiste à détecter le jumeau d'un locuteur dans un ensemble fermé, i.e. l'ensemble des locuteurs (et des enregistrements) est connu. Dans un deuxième temps (cf. *Section 4*), une expérience de Vérification Automatique du Locuteur en milieu ouvert (i.e. nous ne possédons pas d'information sur des locuteurs hormis le locuteur cible) est présentée. Enfin, nous concluons dans la *Section 5* et donnons les limites et les perspectives de l'étude présentée dans ce papier.

2. CORPUS ET OUTILS STATISTIQUES

Cette partie est consacrée à la description de la base de données et aux outils statistiques employés.

2.1. Corpus utilisé

La base de données est composée d'enregistrements de 17 couples de frères et soeurs jumeaux homozygotes (10 couples femmes, 7 couples hommes)[1]. Chacun des locuteurs a lu, entre autre, un passage de "La chèvre de M. Seguin" (Alphonse Daudet), d'une durée comprise entre 40 et 70 secondes selon les locuteurs. Les informations disponibles sur les locuteurs sont rudimentaires : identité, sexe, âge, lieu de naissance, fumeur/non fumeur et pathologie éventuelle liée à la voix. Aucun questionnaire poussé n'a pu être effectué.

2.2. Outils statistiques utilisés

Par la suite, on note i le numéro d'un locuteur dans la base, x_i et X_{x_i}, respectivement un signal appartenant au locuteur i et le modèle du locuteur i appris à l'aide du signal x_i. Nous notons également $\ell(x_i|X_{x_j})$ la vraisemblance de x_i connaissant le modèle X_{x_j}.

Pour un vecteur y_t de dimension d, la distribution gaussienne multidimensionnelle, notée $\mathcal{N}(\mu, \Sigma)$, a une fonction de densité de probabilité $f_{\mu,\Sigma}(y_t)$ de la forme :

$$f_{\mu,\Sigma}(y_t) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma)}} e^{(-\frac{1}{2}(y_t-\mu)^T\Sigma^{-1}(y_t-\mu))} \quad (1)$$

où μ est le vecteur de moyenne de dimension d et Σ la matrice de covariance de dimension $d \times d$ de la distribution. La fonction $\ell(y_t|\mu, \Sigma) = f_{\mu,\Sigma}(y_t)$ est appelée fonction de vraisemblance de la distribution.

Les modèles X_{x_i} utilisés sont des GMM (Modèles de Mélange de Gaussiennes). Un GMM X est une somme pondérée de gaussiennes multivariables (eq 1) défini par le vecteur de paramètres $\theta_X = (c_1, \ldots, c_K, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K)$, où K est le nombre de composantes et c_k le poids de la mixture associée à la $k^{\text{ème}}$ composante contraint par : $c_k \geq 0$ et

[1] Ces enregistrements ont été réalisés dans le cadre de l'émission intitulée "Les Jumeaux : l'expérience inédite", réalisée par W9 Productions et diffusée sur la chaîne télévisée M6. Dans ce cadre, 50 paires de jumeaux ont subi différents tests dans plusieurs domaines : biométrie, physiologie, comportement, empreintes digitales, perception musicale, production vocale... Ces épreuves se sont déroulées sur un plateau télé comportant plus de 200 personnes, jumeaux et professionnels compris, ce qui a rendu difficile les enregistrements sonores spécifiques à la production vocale. Ceux-ci ont été réalisés à l'aide d'un microphone serre-tête unidirectionnel AKG C420, ce qui a permis d'obtenir des données relativement peu bruitées.

$\sum_{i=1}^{K} c_i = 1$.

Soit X le GMM défini ci-dessus, la vraisemblance pour qu'un vecteur de test y_t soit produit par ce mélange de gaussiennes s'exprime de la façon suivante :

$$\ell(y_t|X) = \ell(y_t|\theta_X) = \sum_{k=1}^{K} c_k \, \ell(y_t|\mu_k, \Sigma_k) \qquad (2)$$

Les modèles GMM sont largement utilisés car ils permettent de modéliser un grand nombre de distributions complexes. Pour un signal de parole y formé de n échantillons $y = \{y_1, \ldots, y_n\}$, la vraisemblance de ce signal connaissant le modèle GMM X est donnée par :

$$\ell(y|X) = \Pi_{i=1}^{N} \ell(y_i|X) \qquad (3)$$

où y_i est le $i^{\text{ème}}$ échantillon du signal y.

2.3. Conditions de l'expérimentation

Les productions sonores ont été directement numérisées sur place au format WAV à la fréquence d'échantillonnage de 25 kHz par l'intermédiaire du dispositif EVA [12]. Ces signaux ont été sous-échantillonnés à une fréquence de 16kHz pour l'expérimentation.

La paramétrisation des signaux a été réalisée par la méthode MFCC (Mel Frequency Cepstrum Coefficients). Toutes les 1/100s, le signal est caractérisé par un vecteur composé de 16 coefficients cepstraux et de leurs dérivées respectives. Une normalisation du type CMS (Cepstral Mean Subtraction) est ensuite appliquée à ces paramètres. Compte tenu de la courte durée des enregistrements, il n'y a pas eu de suppression automatique de trames de silence.

3. IDENTIFICATION DU JUMEAU D'UN LOCUTEUR

Cette expérience a pour but d'évaluer la possibilité d'identification du jumeau d'un locuteur par un système de RAL. Elle permet aussi de souligner la manière dont est prise en compte la présence de locuteurs jumeaux dans un corpus.

Les conditions de l'expérience sont les suivantes :
- l'ensemble du signal disponible pour un locuteur est utilisé pour l'apprentissage de son modèle,
- le contexte est celui d'une identification en milieu fermé, où nous cherchons parmi les locuteurs autre que le locuteur cible, l'enregistrement le plus proche du modèle de celui-ci.

Par la suite, on note a le numéro du locuteur cible et b celui de son jumeau à identifier. La vraisemblance d'un locuteur cible a connaissant son modèle est donnée par :

$$\ell(x_a|X_{x_a}) \qquad (4)$$

Pour que le jumeau b d'un locuteur a donné soit correctement identifié, on doit avoir :

$$\ell(x_b|X_{x_a}) > \max_{i \neq (a,b)} \ell(x_i|X_{x_a}) \qquad (5)$$

Le graphique (figure 1) représente, en pourcentage et pour un locuteur donné, la vraisemblance de son jumeau par rapport à la somme des vraisemblances de tous les locuteurs (hormis le locuteur cible).
Soit plus précisément,

$$\frac{\ell(x_b|X_{x_a})}{\sum_{i=1, i\neq a}^{34} \ell(x_i|X_{x_a})} \qquad (6)$$

Le résultat du meilleur locuteur (autre que le jumeau) est également fourni. Les locuteurs sont classés par couple, i.e. les couples (1,2), (3,4), ... sont des jumeaux.

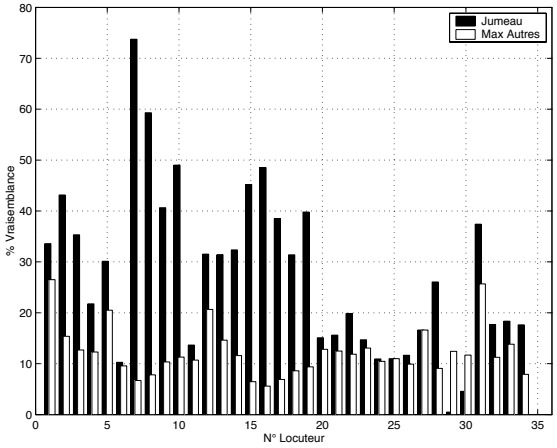

FIG. 1: Vraisemblances relatives du locuteur jumeau et du meilleur imposteur connaissant le modèle du locuteur cible.

Le système reconnaît correctement le jumeau d'un locuteur cible 30 fois sur 34 (soit $\simeq 85\%$ de bonne identification). A quatre reprises, il considère donc comme plus proche du locuteur cible un locuteur autre que son jumeau.

Les informations complémentaires sur les locuteurs nous apprennent que :
- le locuteur 29 était enrhumé pendant l'enregistrement. Le système ne l'a pas identifié comme le jumeau du locuteur 30 et réciproquement,
- le jumeau du locuteur 23, qui venait de subir une opération des cordes vocales, a cependant été correctement identifié,
- en revanche, les jumeaux des locuteurs 25 et 27 qui ne présentaient a priori pas de problèmes particuliers, n'ont pas été identifiés. Par contre la réciproque n'est pas vérifiée (les jumeaux des locuteurs 26 et 28 ont été reconnus).

En conclusion, cette première expérience nous permet d'observer, au sens d'un système de RAL, une ressemblance certaine entre les voix de deux jumeaux.
Cependant sur 34 locuteurs, quatre jumeaux ne sont pas détectés (soit $\simeq 15\%$ d'erreurs), nous confortant dans l'idée que le jumeau d'un locuteur ne constitue pas forcément l'imposteur le plus difficile pour un système de RAL basé sur des méthodes statistiques.

4. VERIFICATION AUTOMATIQUE DU LOCUTEUR

Cette deuxième expérience a pour but de caractériser l'incidence de locuteurs jumeaux sur un système de vérification automatique du locuteur (VAL) en milieu ouvert. Nous réalisons une expérience de vérification classique en milieu ouvert (i.e. aucune information n'est disponible sur les possibles locuteurs imposteurs), basée sur le calcul d'un rapport de vraisemblance (test de type bayésien).
Il est alors nécessaire de construire un modèle de normalisation (représentant l'hypothèse inverse du test bayésien) appelé modèle du monde ou UBM (Universal Background Model) [10].

Le corpus est donc divisé en deux parties, l'une dédiée à la construction de l'UBM et l'autre aux tests.

Pour cette expérience, on définit deux intervalles I_1, I_2 correspondant à la division suivante du corpus :

$$I_1 = [1, 16] \text{ et } I_2 = [16, 34] \quad (7)$$

Les signaux de parole des locuteurs $x_i, i \in I_1$ sont divisés en deux :
- une première partie est d'une durée fixe de 30 secondes, considérée dans cette expérience comme signal d'apprentissage et notée x_{a_A} pour le locuteur a,
- une seconde partie est de durée variable selon la vitesse de prononciation du locuteur. Cette partie est considérée comme signal de test et notée x_{a_T} pour le locuteur a.

Le modèle du monde noté W_1 est appris[2] à partir des signaux x_i avec $i \in I_2$.
Les modèles des locuteurs a, avec $(a \in I_1)$, sont ensuite adaptés à partir du modèle du monde W_1 et des signaux x_{a_A} correspondants[3].

Etant donné un locuteur cible c, x_T un signal de test et les hypothèses H_1 et H_0 suivantes :
H_1 : $\{x_T$ a été prononcé par le locuteur $c\}$
H_0 : $\{x_T$ a été prononcé par un autre locuteur $i \neq c\}$
H_0 représente l'hypothèse inverse de H_1.
La vérification consiste à effectuer le test bayésien ci-dessous :

$$\frac{p(H_1)}{p(H_0)} \geq \lambda \quad (8)$$

où λ est le seuil de décision préalablement fixé.

En VAL, il est largement admis que le rapport ci-dessus peut être estimé par un rapport de vraisemblance (Likelihood Ratio) [11] :

$$LR(x_T, c) = \frac{\ell(x_T|X_{x_{c_A}})}{\ell(x_T|W_1)} \geq \lambda \quad (9)$$

où $\ell(x_{c_T}|X_{x_{c_A}})$ est l'estimation de $p(H_1)$ et $\ell(x_{c_T}|W_1)$ est celle de l'hypothèse inverse $p(H_0)$.

L'expérience est ensuite réitérée en inversant les rôles des deux parties du corpus (l'UBM W_2 est appris sur I_1 et les tests sont menés sur I_2). Nous sommes en présence de deux séries de résultats, que nous fusionnons pour l'analyse.

Trois classes sont définies pour l'expérience :
- une classe Cible, correspondant aux tests pour lesquels le signal à vérifier a été prononcé par le locuteur cible. Il n'y a qu'un seul test par locuteur cible soit 34 tests en tout (nous disposons d'un seul enregistrement par locuteur, servant pour une partie à l'adaptation du modèle du locuteur, et, pour le reste aux tests),
- une classe Jumeau, représentant les tests prononcés par le jumeau du locuteur à vérifier, avec un jumeau par locuteur cible soit 34 tests en tout,
- une classe Imposteurs, représentant les tests qui n'ont pas été prononcés par le locuteur cible ou son jumeau, soit 512 tests en tout (14×16 tests pour la première expérience et 16×18 tests pour la deuxième). Les tests effectués sont en partie croisés Homme/Femme.

[2]Les modèles du monde (UBM) sont des GMM à 128 gaussiennes diagonales, résultant de 20 itérations de l'algorithme EM (Expectation Maximisation) [11].

[3]Les modèles ont été adaptés par la méthode MAP (Maximum A Posteriori)[5][11].

Pour les différentes classes, nous calculons la moyenne, le minimum, le maximum et l'écart-type σ des rapports de vraisemblance définis en (eq 9). Le tableau (table 1) présente les résultats de l'expérience.

Pour calculer un intervalle de confiance sur la moyenne définie ci-dessus, nous utilisons la méthode du *bootstrap*, permettant à partir d'un échantillon et par génération d'une population mère composée d'une multitude d'échantillons de même taille, de calculer un intervalle de confiance pour une statistique donnée.
Précisément, si $LR = \{LR_1, \ldots, LR_n\}$ représente le vecteur des rapports de vraisemblance pour une classe et μ_{LR}, σ_{LR}, respectivement la moyenne et l'écart-type de ce vecteur, LR est considéré comme un échantillon d'une population mère inconnue.
En utilisant un générateur de nombres pseudo-aléatoire, N échantillons notés LR_i, de taille n, sont générés aléatoirement (avec remise) suivant la loi normale $\mathcal{N}(\mu_{LR}, \sigma_{LR})$. Pour chacun de ces échantillons, sa moyenne $\hat{\mu_i}$ est calculée. Toutes ces valeurs sont alors triées pour obtenir : $\hat{\mu}_{(1)} \leq \hat{\mu}_{(2)} \leq \ldots \leq \hat{\mu}_{(N)}$, où $\hat{\mu}_{(k)}$ est le $k^{\text{ème}}$ plus petit des $\hat{\mu_i}$ triés. La méthode du *bootstrap* définit l'intervalle de confiance à $100(1-\alpha)\%$ par :

$$[\hat{\mu}_{(q_1)}, \hat{\mu}_{(q_2)}] \text{ où } \begin{cases} q_1 = \lfloor \frac{N\,\alpha}{2} \rfloor \\ q_2 = N - q_1 + 1 \end{cases} \quad (10)$$

où $\lfloor . \rfloor$ est la partie entière.
Pour cette expérience : $N = 1000$, $\alpha = 0.05$ (l'intervalle de confiance calculé est à 95%), d'où $q_1 = 25$ et $q_2 = 976$. Les intervalles de confiance pour la moyenne des classes sont fournis dans le tableau (table 1).

Le système obtient les performances suivantes pour un seuil global fixé à posteriori :
- pour 0% de fausse acceptation (jumeau du locuteur compris), le système accepte un taux de faux rejet de 53% des locuteurs cibles (soit 18 tests),
- pour 0% de faux rejet, le système a un taux de fausse acceptation de 6% (dans ce cas, il s'agit de deux tests).

Quelques remarques peuvent être faites :
- le système obtient des taux de performances moyens mais, avec un seuil fixé indépendamment du locuteur et sans qu'aucune normalisation du type Znorm ou Tnorm [1] n'ait été réalisée,
- à l'instar de la première expérience, les signaux des jumeaux des locuteurs ont une vraisemblance plus grande que celles des imposteurs,
- les classes Cible et Imposteurs sont clairement dissociées. La vérification ne pose donc pas de problème particulier en présence d'imposteurs classiques,
- les classes Cible et Jumeau présentent un recouvrement induisant toutes les erreurs du système,
- les résultats montrent une forte variation pour les classes Cible ($\sigma \simeq 9$) et Jumeau ($\sigma \simeq 1.5$),
- les intervalles de confiance permettent de relativiser les résultats par rapport au nombre des tests effectués. Pour la classe Imposteurs, le nombre de tests est suffisant pour pouvoir caractériser la distribution de cette classe. En revanche, ce calcul souligne l'incertitude des moyennes sur les deux autres classes.

La présence de locuteurs jumeaux est donc à prendre en compte pour la conception d'un système de vérification automatique du locuteur.

TAB. 1: Rapport de vraisemblance : Moyenne et Intervalle de Confiance à 95% obtenu par la méthode bootstrap, Ecart-type, Minimum et Maximum pour les classes de locuteur Cible, Jumeau et Imposteurs

	Nb Test	**Moyenne**	**[IdC à 95%]**	**Min**	**Max**	σ
Cible	34	11.00	[8.00 , 14.19]	4.56	51.10	8.97
Jumeau	34	2.41	[1.91 , 2.90]	0.73	8.40	1.47
Imposteurs	512	0.82	[0.79 , 0.85]	0.25	2.34	0.36

5. CONCLUSION

Nous avons dans cet article réalisé deux expériences mettant en exergue le comportement d'un système de reconnaissance automatique du locuteur (RAL) lorsqu'il est utilisé sur un corpus contenant des couples de locuteurs jumeaux.

Dans un premier temps, une expérience de détection du jumeau d'un locuteur en milieu fermé est proposée. Nous montrons que des jumeaux ont des voix proches au sens d'un système de RAL (85 % de bonne détection).
La deuxième expérience propose une situation de vérification automatique du locuteur (VAL). Il est clairement apparu difficile de différencier un locuteur de son jumeau. En effet, le système obtient 6% de fausses acceptations pour 0% de faux rejet et 53% de faux rejets pour 0% de fausse acceptation.

Il faut ensuite relativiser cette étude par le manque manifeste de données. En effet, la base de données d'imposteurs est de taille réduite et les tests ont été croisés par genre (580 tests réalisés à comparer aux dizaines de milliers de tests des campagnes d'évaluation internationales du National Institute of Standard and Technology NIST[4] [3]).

Les perspectives de cette étude sont multiples. Il serait intéressant de comparer des jumeaux de même sexe et de même tranche d'âge à l'aide d'un corpus plus conséquent, pour conforter ces résultats. Nous avons toutefois conscience des difficultés de recruter ce type très spécifique de locuteurs.
Pour compléter notre démarche, réalisée à partir de l'étude des coefficients cepstraux (et de leurs dérivées), un système basé sur des critères prosodiques pourrait être implémenté pour évaluer l'influence de la gémellité sur le rythme et/ou l'intonation des locuteurs. Plus généralement, l'étude de paramètres de stabilité de la voix (jitter, shimmer, coefficient de variation de la F0) pourrait permettre de mettre en évidence les aspects physiologiques au niveau de la source laryngée. Enfin, à l'instar de Nolan [9], une analyse de la coarticulation pourrait laisser apparaître des similarités, conséquences de ressemblances anatomiques de l'appareil phonatoire mais aussi des différences dues à une utilisation personnalisée.

Enfin, le champ de cet article pourrait être étendu par l'étude de l'influence de la taille de la base de données sur l'interprétation des résultats obtenus par un système de RAL. La réalisation d'une expérience similaire sur des petits échantillons tirés, aléatoirement ou suivant des critères précis, d'un corpus de taille conséquente, compléterait notre étude. En effet, les résultats et leurs intervalles de confiance calculés dans ce papier renforcent l'idée d'extrême précaution [2] dont il faut faire preuve pour toute application relevant de la vérification automatique du locuteur.

[4] http ://www.nist.gov/speech

RÉFÉRENCES

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification system. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.

[2] J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds, and Y. Magrin-Chagnolleau. Person authentification by voice : A need for caution. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 33–36, Geneve (Suisse), Septembre 2003.

[3] G. R. Doddington. Speaker recognition evaluation methodology – an overview and perspective –. In *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 60–66, Avignon (France), Avril 1998.

[4] C. Fredouille, J.-F. Bonastre, and T. Merlin. AMIRAL : a block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), Janvier/Avril/Juillet 2000.

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

[6] L. Gedda, L. Fiori-Ratti, and G. Bruno. *La voix chez les jumeaux monozygotiques*, volume 12. Folia Phoniatrica, 1960.

[7] M. Homayounpour and G. Chollet. A study of intra- and inter-speaker variability in the voices of twins for speaker verification. In *International Congress of Phonetic Sciences*, 1995.

[8] John L. Locke and L. Mather, Patricia. Genetic factors in the ontogeny of spoken language : Evidence from monozygotic and dizygotic twins. *Journal of Child Language*, 16, 1989.

[9] Francis Nolan and Tomasina Oh. Identical twins, different voices. *Forensic Linguistics*, 3, 1996.

[10] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Speech Communication*, volume 17(1-2), pages 91–108, 1995.

[11] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *European Conference on Speech Communication and Technology (Eurospeech)*, Rhôdes (Grèce), Septembre 1997.

[12] B. Teston, A. Ghio, and B. Galindo. A multisensor data acquisition and processing system for speech production investigation. In *Proceedings of 14th International Congress of Phonetic Sciences*, pages 2251–2254, San Fransisco (USA), 1997.