Voix Propres: Une représentation compacte de locuteurs dans l'espace des modèles

P. Nguyen[1,2], R. Kuhn[1], J.-C. Junqua[1], N. Niedzielski[1], C. Wellekens[2]
[1] Speech Technology Laboratory, Santa Barbara, Californie
[2] Institut Eurécom, Sophia-Antipolis, France
Patrick.Nguyen@eurecom.fr

Résumé: Une nouvelle technique d'adaptation au locuteur pour la reconnaissance de la parole ([3, 4, 5]) est apparue récemment. Son nom, *eigenvoices* ou voix propres, provient de la similarité avec la technique de reconnaissance de visages en image, *eigenfaces*. Dans ce résumé, nous donnons une introduction générale, définissons des domaines d'applicabilité puis développons l'étude de ses performances sur l'adaptation au locuteur pour la reconnaissance de la parole ainsi qu'en reconnaissance du locuteur.

1 Introduction

Dans cette section, nous expliquons le concept dans sa généralité.

1.1 Inspiration: Eigenfaces

L'inspiration initiale, et le nom même de la technique nous vient du génie de l'imagerie. La reconnaissance de visages consiste à reconnaître une image formée de pixels 2D parmi celles présentes dans une base de données. En appliquant les techniques simples de traitement du signal au problème, on se trouve rapidement face à un obstacle de complexité. Ainsi les chercheurs ont compris que la dimensionalité de "l'espace des visages", ie l'espace de variations entre photographies de visages humains est bien plus réduite que la dimensionalité apparente, à savoir celle des images 2D. Comme approximation, on peut considérer un visage comme une combinaison linéaire d'un nombre réduit de composants d'image dits *visages propres* dérivés d'un jeu d'images de références. On applique alors PCA à notre base de données pour obtenir ces visages propres.

Pour effectuer la reconnaissance, il nous suffit juste de projeter l'image dans l'espace des visages et de comparer la position de l'image dans cet espace réduit et les visages connus.

1.2 Voix propres: concept

L'idée directrice est que la dimensionalité des problèmes attaqués est très haute en comparaison à la dimensionalité réelle. Nous faisons donc appel à une technique de réduction de dimensionalité automatique dans l'espace des modèles de locuteurs de façon à faciliter en l'étude.

Concrètement, soit λ l'ensemble des paramètres qui décrivent un modèle. Nous voulons des dépendants du locuteur. Par exemple, dans un système de codage de la parole, ce pourrait être des excitations glottales de notre locuteur en question. Dans un système de reconnaissance du locuteur, nous traiterions les paramètres des Generalized Markov Models (GMMs). Soit D la dimensionalityé de λ. Nous observons la distribution spatiale d'un ensemble de T locuteurs dans $\mathbb{R}^D$. Ensuite nous trouvons une représentation *compacte* de cet espace sous l'assomption de linéarité. Nous postulons donc que les modèles de locuteurs se trouvent dans un sous-espace vectoriel de dimension E, soit:

$$\lambda = \sum_{e=1}^{E} w_e \bar{\lambda}_e, \qquad e = 1, ..., E \tag{1}$$

où $\bar{\lambda}_e$ est appelé une *voix propre* et w_e représente donc la valeur de notre locuteur dans cette direction. L'espace des locuteurs est donc donné par l'ensemble des $\bar{\lambda}_e$ et à chaque locuteur sera associé un vecteur de *charactéristiques* propres $w = [w_1, ..., w_E]^T$.

Par exemple, dans nos expériences, nous avons employé l'analyse en composantes principales (PCA) dans l'espace des paramètres d'un système de reconnaissance de la parole basé sur les chaînes de Markov cachées (HMM). Il s'est avéré que les sources principales de variabilité dans l'espace des modèles observés furent respectivement le sexe du locuteur et le volume de la parole.

2 Voix propres

Cette section vise à explorer plus en détails les voix propres en parole. Nous la divisions en deux parties: d'abord, nous donnons une vue d'ensemble d'un système utilisant les voix propres, et ensuite nous formulons les voix propres dans le contexte des chaînes de Markov cachées ainsi que les estimateurs optimaux correspondants.

2.1 La description du système

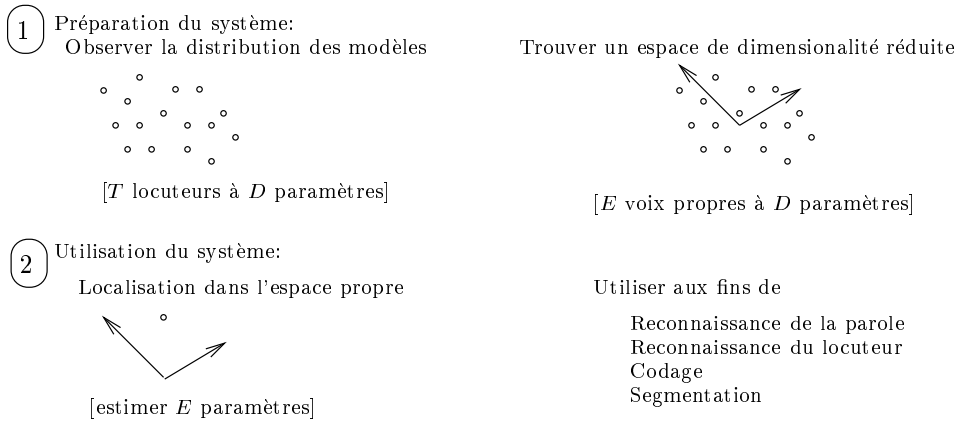
Dans cette sous-section, nous décrivons en général le système de voix propres pour l'adaptation afin de comprendre où se placent les éléments.

Ainsi que le résume la figure 2.1, il y a principalement *deux* phases

1. Préparation du système: cette phase est en général appelée phase d'entraînement dans un système de reconnaissance de parole. Ici on construit les modèles d'information *a priori*. On dit aussi que nous travaillons *hors ligne* car c'est un exercice auquel nous nous prêtons une fois pour toutes.

2. Utilisation du système: ceci correspond a la phase de test dans un système de reconnaissance de la parole. Nous utilisons notre connaissance *a priori* pour effectuer notre tâche au mieux.

Figure 2.1 Vue générale d'un système a voix propres



Dans la phase de préparation, nous observons la densité de probabilité dans l'espace $\mathbb{R}^D$: nous observons un grand nombre T de locuteurs (typiquement 100–500), chacun ayant un vecteur de D paramètres régissant leur modèle. Plus D augmente, et plus le système dépendant du locuteur admet de degrés de liberté pour mieux modéliser la parole, donc D est relativement élevé (de l'ordre de $2000--16000$). Ensuite, nous appliquons un algorithme de réduction de dimensionalité, par exemple PCA, ICA (analyse en composantes indépendantes), ou LDA (analyse linéaire discriminante), pour obtenir des vecteurs de base definissant un sous-espace vectoriel. Il y a E de ces vecteurs, chacun de dimension D et nous les appelons *voix propres*. Nous pouvons aussi appliquer le critère de *maximum de vraisemblance* (ML) dans le contexte de HMMs (voir 2.2.2). Les voix propres approximent l'espace des locuteurs et donc le sous-espace vectoriel engendré est appelé *espace propre* par extension. Ayant construit notre connaissance *a priori* nous pouvons déployer le système.

Dans la phase d'utilisation, alors nous connaissons les voix propres et n'avons plus qu'à localiser un nouveau locuteur dans l'espace propre, donc estimer E paramètres (voir 2.2.1). Dans le cadre de l'adaptation au locuteur ou le codage de la parole, alors on peut reconstruire le modèle à D paramètres correspondant.

2.2 Voix propres et HMMs: estimateurs optimaux

Le modèle très populaire des chaînes de Markov cachées permet de formuler des estimations optimales pour la localisation d'un locuteur, ainsi que pour les voix propres. La théorie des HMMs est bien connue [7]. Vu la nature cachée des chaînes de Markov, nous avons recours à l'algorithme d'espérance-maximisation (EM, voir [1]). Il s'agit de maximiser la vraisemblance d'àpres nos paramètres θ,

$$\hat{\theta} = \arg\max_{\theta} L(O|\theta) \tag{2}$$

où O est l'observation, et $L(\cdot)$ la fonction de vraisemblance. D'après EM, nous sommes réduits à maximiser la fonction auxiliaire $Q(\cdot,\cdot)$ iterativement sous $\hat{\theta}$,

$$Q(\theta,\hat{\theta}) = \mathrm{E}\left[\log L(O,\zeta|\theta)|O,\hat{\theta}\right] \tag{3}$$

avec ζ les données cachées estimées avec θ, l'estimation courante des paramètres. On montre que dans le cas de l'adaptation des moyennes, cette fonction se réduit simplement à:

$$Q_b(\lambda,\hat{\lambda}) = -\frac{1}{2}P(O|\lambda)\sum_{\substack{\text{états } s\\ \text{de } \lambda}}^{S_\lambda}\sum_{\substack{\text{mixt gauss } m\\ \text{de } s}}^{M_s}\sum_{\text{temps } t}^{T}\left\{\gamma_m^{(s)}(t)[n\log(2\pi)+\log|C_m^{(s)}|+h(\mathbf{o}_t,s)]\right\} \tag{4}$$

avec

$$h(\mathbf{o}_t,s) = (\mathbf{o}_t-\hat{\mu}_m^{(s)})^T C_m^{(s)-1}(\mathbf{o}_t-\hat{\mu}_m^{(s)}) \tag{5}$$

et soient

λ	les paramètres de la HMM
$\mathbf{o}_t$	le vecteur d'observation au temps t
$C_m^{(s)-1}$	la précision (inverse de la covariance) pour la gaussienne m de l'état s
$\hat{\mu}_m^{(s)}$	la nouvelle moyenne pour l'état s, composant de mixture m
$\gamma_m^{(s)}(t)$	la P(utilisant $m|\lambda,\mathbf{o}_t$)

2.2.1 Placement optimal dans l'espace des locuteurs: MLED

Il nous faut estimer ici les coefficients propres, ainsi $\theta = w = [w_1,...,w_E]^T$. Les paramètres cachés ζ sont simplement la segmentation inconnue. Si $\bar{\mu}_m^{(s)}(e)$, $e = 1,...,E$ sont les voix propres, alors par définition

$$\hat{\mu}_m^{(s)} = \sum_{e=1}^{E} w_e\bar{\mu}_m^{(s)}(e), \quad e = 1,...,E \tag{6}$$

Pour maximiser la fonction auxiliaire $Q(\cdot,\cdot)$, il faut

$$\frac{\partial}{\partial w_e}Q(\lambda,\hat{\lambda}) = 0 = \sum_{\substack{\text{états } s\\ \text{de } \lambda}}^{S_\lambda}\sum_{\substack{\text{mixt gauss } m\\ \text{de } s}}^{M_s}\sum_{\text{temps } t}^{T}\left\{\frac{\partial}{\partial w_e}\gamma_m^{(s)}(t)h(\mathbf{o}_t,s)\right\}, \quad e = 1,...,E \tag{7}$$

Après quelque arithmétique nous nous trouvons simplement face au système d'équations linéaires suivant:

$$\sum_s\sum_m\sum_t \gamma_m^{(s)}(t)\bar{\mu}_m^{(s)T}(e)C_m^{(s)-1}\mathbf{o}_t = \sum_s\sum_m\sum_t \gamma_m^{(s)}(t)\sum_{j=1}^{E} w_j\bar{\mu}_m^{(s)T}(j)C_m^{(s)-1}\bar{\mu}_m^{(s)}(e), \tag{8}$$

en w_e, $e = 1,...,E$, ce qui revient a inverser une matrice $E\times E$.

2.2.2 Approximation optimale de l'espace des locuteurs: MLES

Nous dérivons maintenant une méthode simple pour trouver l'espace propre. Nommons cet espace $M = [\bar{\mu}_m^{(s)T}(1), ..., \bar{\mu}_m^{(s)T}(E)]^T$. Cette méthode est appelée espace propre à maximum de vraisemblance (MLES). Elle a plusieurs avantages. D'abord, PCA exige beaucoup de mémoire. Ensuite, MLES ne minimise pas une mesure de divergence entre gaussiennes du même modèle, même état, ce qui en particulier appelle à un alignement entre gaussiennes de locuteur à locuteur. MLES permet d'exprimer des connaissances *a priori* dans l'espace propre.

Pour résoudre le problème via EM, nous étendons ζ aux w. Il s'agit de trouver $\theta = M$. Il vient:

$$\hat{M} = \arg\max_M \sum_{q=1}^{T} \int \log L(O, w|M) P_0(w, q)\, dw \tag{9}$$

où $P_0(w, q)$ porte l'information *a priori* sur le locuteur q (eg la probabilité d'observer une personne de tel ou tel sexe, dialecte, niveau d'études, etc, conjointement avec ces valeurs propres). Cette probabilité est essentielle pour les bases de données non-équilibrées (eg trop de mâles). Supposons encore pour la simplicité de l'exposé que $P_0(w, q) = P_0(q)\Pi_{k=1}^{E} P_0(w_k|q)$. Un exemple d'une telle fonction serait:

$$P_0(w_k|q) = \begin{cases} 1 & \text{si } w_k > 0 \text{ et le locuteur } q \text{ est masculin} \\ 1 & \text{si } w_k < 0 \text{ et le locuteur } q \text{ est féminin} \\ 0 & \text{ailleurs} \end{cases}$$

Les voix propres initiales peuvent être obtenues par PCA, ICA, ou LDA. Quand aucune connaissance *a priori* n'est disponible, alors on utilise MLED décrit plus haut pour remplacer l'intégration dans (9) par l'operateur maximum. On dérive les formules de réestimation de façon tout-à-fait similaire à l'algorithme de Baum-Welch:

$$\bar{\mu}_e^{(m)} = \frac{\sum_q L_q w_q^{(e)} \sum_t \gamma_m(t) \left\{ \mathbf{o}_t - \tilde{\mu}_q^{(m)}(e) \right\}}{\sum_q L_q \left(w_q^{(e)} \right)^2 \sum_t \gamma_m(t)} \tag{10}$$

où q, m, e representent un locuteur, une distribution du modèle, et une voix propre respectivement. De plus, L_q est la probabilité *a posteriori* des échantillons de parole $O^{(q)}$ du locuteur q, à savoir $L_q = L(O^{(q)}|w_q^{(e)})p(w_q^{(e)})$. A nouveau, $\gamma_m(t)$ la probabilité *a posteriori* d'observer la distribution m. L'estimation courante de la $e^{\text{ième}}$ valeur propre du locuteur q est $w_q^{(e)}$. On définit finalement le complément de la voix propre $\tilde{\mu}_q^{(m)}$

$$\tilde{\mu}_q^{(m)}(e) = \sum_{k=1, k\neq e}^{E} w_q^{(k)} \bar{\mu}_k^{(m)}, \qquad e = 1, ..., E \tag{11}$$

Ainsi l'algorithme que nous proposons ressemble à une passe de réestimation de Baum-Welch:

1. Pour chaque locuteur

   - Estimer les valeurs propres optimales par MLED (plusieurs iterations)
   - accumuler pour MLES en prenant garde d'utiliser $P_0(\cdot)$

2. Réeestimer les voix propres et recommencer

En comparant avec l'entraînement du modèle indépendant du locuteur (SI), alors nous remarquons qu'en termes de mémoire nous utilisons E accumulateurs de dimension D au lieu d'un seul, et qu'une itération prend approximativement autant de fois plus de temps qu'il y a d'itérations pour trouver l'estimateur MLED. Finalement, notons que PCA donne l'estimateur des moindres carrés alors que MLES nous offre l'estimateur sous le critère ML ou MAP.

3 Applications

Les applications en technologies de la parole sont légion. Partout où des systèmes dépendants du locuteur peuvent être mis en place, alors le concept des voix propres permet de réduire efficacement la dimensionalité du problème et donc d'en faciliter sa résolution. Par exemple le codage de la parole peut grandement tirer parti d'une description compacte du locuteur, car par définition peu de paramètres vont être transmis. Dans le reste de cette section nous décrivons deux applications pratiques basées sur des HMMs.

3.1 Adaptation au locuteurs

L'adaptation au locuteur se réduit à observer des locuteurs et à en déduire un espace propre. Ensuite, nous utilisons le critère bien connu de maximum de vraisemblance (ML) [3] afin de localiser un nouveau locuteur. Ceci est avantageux par rapport aux techniques d'adaptation au locuteur connues telles que la régression linéaire au maximum de vraisemblance (MLLR) et celle dite de maximum *a posteriori*. L'application des voix propres réduit le nombre de paramètres à estimer et donc est propice à une adaptation rapide dans le sens où ces paramètres sont estimés de façon plus robuste. Aussi, la complexité de la méthode est faible par rapport à MLLR.

3.2 Reconnaissance du locuteur

La représentation intuitive d'un locuteur par sa localisation dans l'espace des locuteurs présume donc que plus des locuteurs seront dissimilaires, plus la distance Euclidienne dans cet espace s'en trouve grandie. Ainsi, nous trouvons les pas suivants pour construire un système de reconnaissance du locuteur

1. Construction de l'espace des locuteurs

   (a) Enrôlement: on construit un modèle dépendant du locuteur (dimension D) selon les enregistrements de sa voix. Il y a T locuteurs.

Methode	$E = 5$	$E = 10$	$E = 20$	$E = 50$
PCA	60.67	60.58	61.29	61.56
MLES($E = 10$)	62.53	65.10	-	-
MLES($E = 20$)	63.06	65.01	65.37	-
MLES($E = 50$)	61.74	63.77	64.84	66.96

Table 1: Espace propre à maximum de vraisemblance (MLES) vs PCA

(b) Reduction de dimensionalité: nous trouvons les $\bar{\lambda}_e, e = 1...E$ décrivant l'espace en utilisant PCA.

2. Modèles de référence: nous n'avons plus besoin de tous les modèles pour la suite des opérations, mais uniquement de leur projections w_e dans l'espace des locuteurs. Nous réduisons donc le système de $D \times T$ à $D \times E$ et l'espace lui-même $E \times D$.

3. Reconnaissance du locuteur: pour un locuteur nouveau présenté, nous le localisons dans l'espace des locuteurs. Le modèle de référence le plus proche est donc l'identité présumée du locuteur.

Afin de comparer la proximité dans l'espace des locuteurs, nous utilisons l'angle entre les vecteurs de valeurs propres. Soient deux locuteurs q et r définis par w_q et w_r. Alors la similarité est:

$$\delta(r, q) = \arg\cos\left(\frac{w_r^T w_q}{\sqrt{w_r^T w_r \cdot w_q^T w_q}}\right) \tag{12}$$

Pour un système de vérification du locuteur, on accepte deux locuteurs comme étant les identiques suivant un seuil heuristique K si $\delta(r, q) > K$, et on rejette sinon.

4 Expérimentations

Nos expériences furent effectuées sur la base de données TIMIT. Celle-ci est composée de 462 locuteurs pour l'entraînement, chacun prononçant 8 phrases de 2-7 secondes. Nous avons utilisé 18 coefficients PLP [2] et échantillonné a 16 kHz, 48 modèles indépendants du contexte, à 3 états et 16 distributions par état. La base de test est formée par 30 et 150 locuteurs pour l'adaptation à et la reconnaissance du locuteur respectivement, chacun prononçant une phrase. La table 1 résume les resultats pour différentes dimensions $E = 5, 10, 20, 50$ de l'espace propre, entraîné pour $E = 10, 20, 50$ en taux de reconnaissance de phonème. Le système sans adaptation (SI) donne 60.94%. Cette méthode dépasse l'adaptation de l'état de l'art, MLLR, combinée avec MAP, qui se contente d'un modeste 59.64% (pas assez de données pour s'adapter), donc ces résultats sont excellents.

Pour la reconnaissance du locuteur, nous choisissons $E = 20$. L'identification consiste à choisir quel est le bon locuteur, soit donné une phrase. La vérification

consiste à rejeter ou accepter un locuteur se présentant donnant une identité. On peut donc faire deux type d'erreurs, celle de fausse réjection (FR) où le locuteur est le bon mais est rejeté par le système, et celle de fausse acceptation (FA), où un imposteur se présente avec une fausse identité et est accepté par le système. On définit alors le taux d'erreurs égales (EER) pour le pourcentage pour lequel FA=FR, en variant le seuil K. Dans nos expériences, nous avons obtenu 100% d'identification correcte et 2% d'EER lorsque tous les locuteurs jouaient comme imposteurs. Ce sont des résultats modestes mais encourageants.

5 Conclusion

Dans ce papier, nous décrivons le concept des voix propres comme étant une manière de représenter de façon compacte des locuteurs dans le domaine des modèles. Nous montrons aussi des applications en reconnaissance du locuteur et reconnaissance de la parole, avec une spécialisation pour les HMMs. Grâce à la dimension réduite du modèle de locuteur, nous améliorons la robustesse avec laquelle les paramètres sont estimés, donnant voie à des applications jusqu'alors difficiles. Nous donnons dans les détails les estimateurs optimaux pour HMMs, et présentons de rapides incursions dans les domaines de l'adaptation au et de la reconaissance du locteur.

Ainsi nous montrons comment avoir une représentation compacte optimale des modèles et comment l'utiliser dans différents domaines.

References

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum-Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society B*, pages 1–38, 1977.

[2] Hynek Hermansky. Perceptual linear predicitve (PLP) analysis of speech. *Journal of the Acoustic Society of America*, 87(4):1738–1752, April 1990.

[3] R. Kuhn, P. Nguyen, J.-C. Junqua, R. Boman, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Fast Speaker Adaptation in Eigenvoice Space. *ICASSP*, 1999.

[4] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. *ICSLP*, 1998.

[5] Roland Kuhn, Patrick Nguyen, Jean-Claude Junqua, Lloyd Goldwasser, Nancy Niedzielski, Steven Fincke, and Kenneth Field. Eigenfaces and eigenvoices: dimensionality reduction for specialized pattern recognition. *MMSP*, 1998.

[6] P. Nguyen, C. Wellekens, and J.-C. Junqua. Maxmimum Likelihood EigenSpace and MLLR for speech recognition in Noisy Environments. *Eurospeech*, 1999. To appear.

[7] Rabiner and Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1994. ISBN 0-13-015157-2.