Amélioration des taux de reconnaissance par filtrage de données [1]

Chris J. Wellekens
Institut EURÉCOM[2] – Sophia-Antipolis
Christian.Wellekens@eurecom.fr

Résumé: De nombreux résultats récents mettent en évidence une amélioration des taux de reconnaissance de la parole lorsqu'on applique un filtrage non récursif aux trajectoires des vecteurs de caractéristiques. Cet article présente une nouvelle approche où les caractéristiques des filtres sont entrainées en même temps que les paramètres des modèles de mots et qui conduit dans des premiers tests à une amélioration des taux de reconnaissance. Les formules de réestimation des fréquences de coupure des filtres sont déduites ainsi que les coefficients de leur réponse impulsionnelle dans un cas plus général.

1 Introduction

La reconnaissance automatique de la parole repose sur la comparaison de mots prononcés avec des modèles de sous-unités lexicales ou phonétiques. Les modèles les plus utilisés sont les modèles de Markov cachés (HMM) entrainés qui ont été l'objet de nombreuses sophistications dans les dernières années afin d'une part d'augmenter leur vraisemblance vis à vis de la parole réelle et d'autre part de ne pas accroitre de façon excessive le nombre de paramètres requis à leur représentation.

Cependant, l'information pertinente est dissimulée au sein de nombreuses données inutiles pour le reconnaissance de parole. C'est pourquoi avant d'entrainer les modèles, il est indispensable de prétraiter les données. La fréquence fondamentale de l'onde glottique (pitch) ainsi que les déphages entre composantes du spectre sont éliminées. Une analyse harmonique est alors pratiquée (prédiction linéaire, analyse cepstrale, spectres lissés ou bancs de filtres) qui ramène les données dans le domaine purement fréquentiel.

La non-stationarité de la parole impose une analyse sur fenêtres temporelles. La longueur des fenêtres est typiquement de l'ordre de 30ms décalées dans le temps de 10 ms (c'est à dire avec un fort recouvrement): ces valeurs sont consistantes avec l'inertie de l'appareil articulatoire.

Très tôt [1-2], les chercheurs ont observé le rôle important de la dynamique c'est à dire de l'évolution temporelle des caractérisques dans la qualité de la reconnaissance (vitesse (Δ)et de l'accélération ($\Delta\Delta$)). La dimension de l'espace de représentation s'en trouve accrue de même que le temps de calcul et le volume de données requis pour l'entrainement.

[1]This research is an unpublished contribution to the Large Vocabulary Speech Recognition Summer Workshop 1997, CLSP, The Johns Hopkins University, Baltimore. Special thanks are due to H.Hermansky for fruitful discussions

[2]Institut Eurécom est partiellement financé par Ascom, Cegetel, France Telecom, Hitachi, IBM, Motorola, Swisscom,Texas Instruments, Thomson CSF

Des modèles de Markov prédictifs ont également été proposés où la probabilité (ou la distance) d'émission est définie en termes de coefficients de prédiction c'est à dire en tenant compte d'un ou de plusieurs vecteurs précédents, [3-4] entre autres.

Cependant l'accroissement du nombre de paramètres des modèles exige simultanément plus de données pour l'entrainement. Ainsi l'analyse en composantes principales (PCA) a été suggérée pour réduire la dimension de l'espace de représentation en ne gardant que des composantes indépendantes significatives. Cependant comme elle s'applique à l'ensemble des données sans tenir compte de leur appartenance à une classe, elle n'offre que peu de discrimination. L'analyse discriminante linéaire (LDA) proposée depuis longtemps et plus récemment en [5-6] fait usage de la connaissance a priori de la segmentation phonétique pour accroitre la classification entre classes.

Ainsi, l'information associée à une trame de 10ms est extraite d'une fenêtre contextuelle plus large et peut être considérée comme le résultat du filtrage des trajectoires des caractéristiques. Récemment, des coefficients discriminants obtenus par LDA ont été considérés comme des coefficients filtrés [5]: des filtres résultants de l'analyse discriminante ont été analysés et leur comportement est proche des filtres dérivateurs créant les coefficients de vitesse et d'accélération.

Cet article présente les formules de réestimation des paramètres des modèles de Markov et d'un filtre qui traite les caractéristiques du signal et qui est entrainé simultanément avec les HMM afin d'accroitre la vraisemblance (likelihood) sur la base de données.

En section 2, le filtre qui sera appliqué aux trajectoires des vecteurs caractéristiques est décrit tandis que la section 3 est une discussion sur le critère d'entrainement modifié pour tenir compte du filtre.

Les paramètres du filtre sont entrainés en même temps que ceux des HMM. En conséquence, non seulement les paramètres des modèles mais aussi les données sont modifiées durant l'entrainement. On pourrait émettre la critique que le prétraitement des données afin d'accroitre la vraisemblance (likelihood) n'est pas la meilleure façon d'accroitre le taux de reconnaissance: en effet, inclure les fréquences de coupure du filtre dans l'entrainement revient à modifier les données afin qu'elles soient plus proches des modèles Markoviens utilisés. Mais on peut considérer le filtrage comme une partie du modèle et toute tentative d'utiliser l'erreur de prédiction ou toute autre combinaison de vecteurs successifs poursuit un but équivalent.

Un filtre unique commun à toutes les caractéristiques est entrainé. Plus généralement des filtres différents pour chaque caractéristique peuvent être utilisés et même des filtres différents à la fois pour chaque état et chaque caractéristique.

Le filtrage peut conduire à une réduction du flux de données: en effet si la bande passante du flux de caractéristiques est suffisamment réduite, on peut envisager une décimation et donc accélérer le processus de reconnaissance.

2 Description du filtre

Les spécifications d'un filtre s'expriment en temps et/ou en fréquence.

Nous utilisons la réponse impulsionnelle tronquée d'un filtre idéal:

$$h_p(\omega_u) = \frac{\sin(\omega_u p)}{\pi p} \qquad\qquad p \in [-P, \ldots, P]$$

où ω_u est la fréquence de coupure d'un passe-bas de longueur $2P+1$. UN seul paramètre décrit l'ensemble du comportement du filtre.

Si la fréquence de coupure est égale à la fréquence de Nyquist (soit 50 Hz pour un débit de trames (fenêtres) de 10 ms), la réponse impulsionnelle ne possède qu'un seul échantillon non-nul à l'instant $k = 0$ et ceci correspond à l'absence de filtrage. Un effet semblable est obtenu si le filtre est de longueur 1 ($P = 0$).

Dans la suite, les vecteurs acoustiques filtrés seront notés $x^t = (\xi_1, \ldots, \xi_d)$ tandis que les vecteurs originaux seront désignés par $z^t = (\zeta_1, \ldots, \zeta_d)$. Ainsi ζ_j est la j-ème composante du vecteur acoustique z et en appelant $\zeta^{(p)}$ la composante d'un vecteur $z^{(p)}$ situé p trames après z, la version filtrée de ζ est

$$\xi_j = \sum_{p=-P}^{P} h_p(\omega_u)\zeta_j^{(p)}. \tag{1}$$

Toutes les composantes de tous les vecteurs sont modifiées par cette formule et l'entrainement des paramètres des HMM utilisera ces nouveaux vecteurs.

Un problème important est la normalisation de la réponse. La puissance de la réponse est

$$\mathcal{P} = \sum_{p=-P}^{P} h_p^2$$

et dépend bien sur de la fréquence de coupure ω_u. En remplaçant h_p by $h_p/\sqrt{\mathcal{P}}$ on obtient la réponse impulsionnelle normalisée.

L'extension au cas d'un filtre passe-bande s'obtient en remarquant que sa réponse impulsionnelle est la différence entre celles de deux filtres passe-bas avec des coupures respectives ω_l et ω_u et dépend bien sûr de ces deux seuls paramètres.

$$h_p = \frac{\sin(\omega_u p) - \sin(\omega_l p)}{\pi p} = \frac{2}{\pi p}\cos(\sigma p)\sin(\delta p) \quad p \in [-P, \ldots, P]$$

avec la fréquence centrale $\sigma = \frac{\omega_u + \omega_l}{2}$ et la largeur de bande $2\delta = \omega_u - \omega_l$.

Le filtre passe-haut est obtenu comme un filtre passe-bande avec une fréquence supérieure égale à la fréquence de Nyquist ($\omega_u = \pi$).

3 Les HMM et leur algorithme d'entrainement

L'algorithme de Viterbi est utilisé dans cet article pour l'entrainememt comme pour la reconnaissance. Le meilleur chemin fournit une partition de la base de données de sorte que chaque vecteur acoustique est associé à un état particulier. La vraisemblance (likelihood) de l'ensemble d'entrainement est

$$L = \prod_j \prod_{x \in \mathcal{Q}_j} p(x|q_j) P_t$$

où P_t est le produit de toutes les probabilités de transition associées à ce chemin; $\mathcal{Q}_j$ est l'ensemble des vecteurs associés à l'état q_j; le produit en j couvre l'ensemble des états indépendants des modèles et $p(x|q_j)$ est la densité de probabilité (probability density fuction pdf) associée à l'état q_j.

Nous nous limiterons dans la suite à des pdf's monogaussiennes (μ_i and Σ_i designent le vecteur moyen et la matrice de covariance de l'état q_i) puisque notre objectif est une étude de faisabilité.

Si toutes les matrices Σ sont supposées diagonales, la log-vraisemblance Λ c'est à dire $-\log(L)$ est

$$\Lambda = 1/2 \sum_j \sum_{x \in \mathcal{Q}_j} \sum_{k=0}^{d} \left(\frac{\xi_k - \mu_{jk}}{\sigma_{jk}} \right)^2 + \sum_j \frac{n_j}{2} \log((2\pi)^d |\Sigma_j|) - \log(P_t) \tag{2}$$

où n_j est le nombre de vecteurs de $\mathcal{Q}_j$.

La contribution P_t est indépendante de celle des états et peut être traitée séparement sans perte de généralité. Les estimateurs de m_j et de Σ_j obtenus en annulant les dérivées de Λ respectivement par rapport à m_j et Σ_j sont

$$\hat{m}_j = \frac{1}{n_j} \sum_{x \in \mathcal{Q}_j} x \tag{3}$$

et

$$\hat{\Sigma}_j = \frac{1}{n_j} \sum_{x \in \mathcal{Q}_j} (x - m_j)(x - m_j)^t. \tag{4}$$

Il est important de remarquer que si tous les vecteurs sont multipliés par un facteur commun K, les déterminants $|\Sigma_j|$ sont multipliés par K^2 fournissant des termes additionnels à Λ: ceci montre que Λ dépend de l'échelle. La meilleure façon d'éviter cette dépendance d'échelle est de contraindre $\mathcal{P} = 1$ et donc de modifier la log-vraisemblance par un terme Lagrangien.

Ainsi la condition d'optimalité discutée en section 4 dépend du multiplicateur de Lagrange et donc aussi des fréquences de coupure.

Cependant comme on le verra à la section suivante, ni les fréquences de coupure ni le multiplicateur de Lagrange ne peuvent être explicités mais ils résultent d'un processus itératif. Afin de pallier cet inconvénient, les fréquences de coupure sont calculées sans contrainte sur la puissance mais pour éviter une décroissance non-significative de Λ due à ce gain, la réponse impulsionnelle est renormalisée à chaque itération.

4 Réestimation des fréquences de coupure

Afin d'obtenir les formules de réestimation des fréquences de coupure ω_u et ω_l, on annule les dérivées de Λ par rapport à ces variables. Les formules de réestimation sont non-linéaires et dans le cas passe-bande les deux fréquences sont couplées dans un système d'équations non-linéaires.

4.1 Filtre passe-bas

Evidemment, tous les vecteurs dépendent de la fréquence de coupure par (1). En conséquence, il en va de même pour les vecteurs moyens et pour les matrices de covariance par (3-4). Les paramètres à la k-ème itération sont notés $m_j^{<k>}, \Sigma_j^{<k>}, \omega_u^{<k>}$.

Dans un entrainement conventionnel sans pré-filtrage, la différentielle de Λ

$$d\Lambda = \sum_j \left(\frac{\partial \Lambda}{\partial m_j} dm_j + \frac{\partial \Lambda}{\partial \Sigma_j} d\Sigma_j \right)$$

doit être annulée: ce qu'on obtient en annulant toutes les dérivées partielles et les formules (3)(4) s'ensuivent.

Considérons à présent le cas avec filtrage. La dérivée totale de Λ par rapport à ω_u est

$$\frac{d\Lambda}{d\omega_u} = \sum_j \left(\frac{\partial \Lambda}{\partial m_j} \ \frac{dm_j}{d\omega_u} + \frac{\partial \Lambda}{\partial \Sigma_j} \ \frac{d\Sigma_j}{d\omega_u} \right) + \sum_{all \ x} \frac{\partial \Lambda}{\partial x} \ \frac{dx}{d\omega_u}. \tag{5}$$

Puisque les dérivées partielles par rapport à m et Σ entre crochets dans (5) s'annulent grace au choix des estimateurs faits précédemment (3-4), la dérivée totale de Λ par rapport à ω_u s'annulera si

$$\sum_{all \ x} \frac{\partial \Lambda}{\partial x} \ \frac{dx}{d\omega_u} = 0. \tag{6}$$

Faisant usage de (1-2), cette expression devient sous l'hypothèse que tous les Σ_j sont des matrices diagonales

$$\sum_j \sum_{\xi \in \mathcal{Q}_j} \sum_{k=0}^{d} \frac{\xi_k - \mu_{jk}}{\sigma_{jk}^2} \frac{d\xi_k}{d\omega_u} = 0 \tag{7}$$

ou par (1),

$$\sum_{p=-P}^{P} \sum_{q=-P}^{P} A_{pq} \cos(\omega_u p) \frac{\sin(\omega_u q)}{\pi q} = \sum_{p=-P}^{P} A_p \cos(\omega_u p) \tag{8}$$

où

$$A_p = \textstyle\sum_j \sum_{\xi \in \mathcal{Q}_j} \sum_{k=0}^{d} \frac{1}{\sigma_{jk}^2} \mu_{jk} \zeta_k^{(p)} \quad \text{et} \quad A_{pq} = \sum_j \sum_{\xi \in \mathcal{Q}_j} \sum_{k=0}^{d} \frac{1}{\sigma_{jk}^2} \zeta_k^{(p)} \zeta_k^{(q)}.$$

On vérifie aisément que $A_{pq} = A_{qp}$. Les coefficients A_{pq} and A_p contiennent la statistique collectée durant le parcours inverse du chemin optimal (backtracking) dans l'ensemble des phrases de la base d'entrainement.

En explicitant les termes pour lesquels $q = 0$, on trouve:

$$\frac{\omega_u}{\pi} \sum_{p=-P}^{P} A_{p0} \cos(\omega_u p) = \sum_{p=-P}^{P} A_p \cos(\omega_u p) - \sum_{p=-P}^{P} \sum_{q=-P; q \neq 0}^{P} A_{pq} \cos(\omega_u p) \frac{\sin(\omega_u q)}{\pi q} \tag{9}$$

qui est formulation de l'équation du type point fixe $\omega_u = f(\omega_u)$. Cette équation de type point fixe peut être résolue itérativement $\omega_u^{<k+1>} = f(\omega_u^{<k>})$ à chaque itération de l'agorithme Viterbi. La réponse sera simplement normalisée en puissance à chaque itération afin de garantir la contrainte de puissance.

La solution n'est pas unique et le signe de la dérivée seconde de Λ doit être vérifié pour garantir un minimum.

On peut faire remarquer que la solution doit se trouver dans l'intervalle $[-\pi, \pi]$. Cependant, ω_u est uniquement utilisé dans (1). Clairement h_p est une fonction périodique de ω_u qui est définie modulo 2π.

Pour accroitre la discrimination entre phonèmes, différentes caratéristiques peuvent être utilisées pour le calcul des probabilités d'émission associées à l'état ou aux états d'un phonème.

La transformation des caractéristiques peut être vue comme une partie de la description par HMM et conduit à des définitions spécialisées des probabilités locales au même titre que les distributions Gaussiennes.

Nous faisons ci-dessous l'hypothèse que chaque composante pour chaque état est filtrée par un filtre dédié. Les fréquences de coupure sont maintenant notées ω_{kju}. A nouveau, l'équation (7) est cruciale et devient:

$$\sum_{\xi \in \mathcal{Q}_j} \frac{\xi_k - \mu_{jk}}{\sigma_{jk}^2} \frac{d\xi_k}{d\omega_{kju}} = 0. \tag{10}$$

La définition des paramètres A est à présent:

$$A_{pjk} = \textstyle\sum_{\xi \in \mathcal{Q}_j} \frac{1}{\sigma_{jk}^2} \mu_{jk} \zeta_k^{(p)} \quad \text{et} \quad A_{pqjk} = \textstyle\sum_{\xi \in \mathcal{Q}_j} \frac{1}{\sigma_{jk}^2} \zeta_k^{(p)} \zeta_k^{(q)}.$$

Le nombre de fréquences de coupure à estimer est dS où S désigne le nombre de différents états.

Il est cependant très facile de contraindre tous les états d'un même phonème à partager les mêmes filtres.

4.2 Filtrage passe-bande

Etendons à présent les résultats de la section précédente au cas des filtres passe-bande. La dérivée totale de Λ par rapport à ω_l a une forme semblable à (5). Un argument semblable conduit à une condition supplémentaire

$$\sum_{all\ x} \frac{\partial \Lambda}{\partial x} \frac{dx}{d\omega_l} = 0$$

ou

$$\sum_j \sum_{\xi \in \mathcal{Q}_j} \sum_{k=0}^{d} \frac{\xi_k - \mu_{jk}}{\sigma_{jk}^2} \frac{d\xi_k}{d\omega_l} = 0 \tag{11}$$

où la contrainte de puissance a également été négligée.

Prenant en compte la définition de la réponse impulsionnelle du passe-bande idéal, (7) et (11) fournissent

$$\sum_{p=-P}^{P} \sum_{q=-P}^{P} A_{pq} \cos(\omega_u p) \frac{\sin(\omega_u q) - \sin(\omega_l q)}{\pi q} = \sum_{p=-P}^{P} A_p \cos(\omega_u p) \tag{12}$$

$$\sum_{p=-P}^{P} \sum_{q=-P}^{P} A_{pq} \cos(\omega_l p) \frac{\sin(\omega_u q) - \sin(\omega_l q)}{\pi q} = \sum_{p=-P}^{P} A_p \cos(\omega_l p) \tag{13}$$

avec des définitions identiques des coefficients A. Ecrivons (12) légèrement différemment:

$$\sum_{p=-P}^{P}\sum_{q=-P}^{P} A_{pq}\cos(\omega_u p)\frac{\sin(\omega_l q)}{\pi q} = -\sum_{p=-P}^{P}\left(A_p - \sum_{q=-P}^{P} A_{pq}\frac{\sin(\omega_u q)}{\pi q}\right)\cos(\omega_u p)$$

et (13)

$$\sum_{p=-P}^{P}\sum_{q=-P}^{P} A_{pq}\cos(\omega_l p)\frac{\sin(\omega_u q)}{\pi q} = \sum_{p=-P}^{P}\left(A_p + \sum_{q=-P}^{P} A_{pq}\frac{\sin(\omega_l q)}{\pi q}\right)\cos(\omega_l p)$$

Isolant les termes en $q = 0$ dans les membres de gauche des deux équations fournit les estimateurs point fixe pour les deux fréquences de coupure. Ces équations seront résolues itérativement comme suit:

$$\frac{\omega_u^{<k+1>}}{\pi} = \frac{\omega_l^{<k>}}{\pi} + \left(\sum_{p=-P}^{P}\cos(\omega_u^{<k>}p)\left(A_p - \sum_{q=-P;q\neq 0}^{P} A_{pq}\,\frac{\sin(\omega_u^{<k>}q) - \sin(\omega_l^{<k>}q)}{\pi q}\right)\right) \left(\sum_{p=-P}^{P} A_{p0}\cos(\omega_u^{<k>}p)\right)^{-1} \tag{14}$$

et

$$\frac{\omega_l^{<k+1>}}{\pi} = \frac{\omega_u^{<k>}}{\pi} - \left(\sum_{p=-P}^{P}\cos(\omega_l^{<k>}p)\left(A_p + \sum_{q=-P;q\neq 0}^{P} A_{pq}\,\frac{\sin(\omega_l^{<k>}q) - \sin(\omega_u^{<k>}q)}{\pi q}\right)\right) \left(\sum_{p=-P}^{P} A_{p0}\cos(\omega_l^{<k>}p)\right)^{-1} \tag{15}$$

On voit immédiatement que (14) devient (9) lorsque $\omega_l = 0$.

4.3 Filtre général défini par sa réponse impulsionnelle

Pour terminer, disons quelques mots d'une autre approche d'optimalisation du filtre. Supposons que le filtre recherché ait une réponse impulsionnelle $h_t \quad \forall t \in [-P,\ldots,P]$ et que ce sont ces coefficients que nous cherchons à optimiser. Comme en (5) et (6), nous trouvons une condition:

$$\sum_{all\ x}\frac{\partial\Lambda}{\partial x}\ \frac{dx}{dh_t} = 0 \quad \forall t \in [-P,\ldots,P] \tag{16}$$

où les composantes de x sont à nouveau définies comme en (1) mais où les coefficients h sont les paramètres libres et ne dépendent plus des fréquences de coupure. En utilisant (1), (16) devient:

$$\sum_{p=-P}^{P} h_p \sum_{\xi \in \mathcal{Q}_j} \zeta_k^{(t)} \zeta_k^{(p)} = \mu \sum_{\xi \in \mathcal{Q}_j} \zeta_k^{(t)} \quad \forall t \in [-P, \ldots, P].$$

Cette expression doit être valable pour tout t et la solution de ce système linéaire fournit le filtre optimal. Le nombre de paramètres à déterminer est dans ce cas $(2P+1)dS$.

Il est intéressant de remarquer que la contrainte de puissance discutée en section 3 peut être prise en compte ici car la solution est celle d'un système linéaire.

5 Expériences

Des premières expériences sur une petite base de données dépendante du locuteur ont montré qu'un filtre de longueur 21 et de fréquence de coupure non-optimisée de 20Hz conduit à 35% d'erreurs contre 37% sans prefiltrage des données pour un décodeur phonétique sans grammaire. Ce taux d'erreur tombe à 33% si 50% de sous-échantillonage est appliqué et à 30% si des pénalités d'entrée dans un nouveau phonème sont réglées. Les filtres optimisés n'ont pas encore fournis de résultats significatifs sur cette base de données trop réduite. Des expériences sont en cours sur TIMIT.

6 Références

1. S.Furui, "Speaker independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. ASSP,*,vol.34, nr 1, pp.52-59, 1986.

2. C.J. Wellekens, "Explicit Time Correlation in Hidden Markov Models for Speech Recognition," *Proc. ICASSP-87*, vol. 1, pp. 384-386, Dallas, April 1987.

3. E.Levin, " Speech recognition using hidden control neural network architecture," *Proc. ICASSP-90*, pp. 433-436, Albuquerque (NM), 1990

4. F.Freitag, *An Application of Predictive Neural Networks to Speech recognition*, Tesi Doctoral, UPC, Barcelona, May 1998

5. S.van Vuuren, H.Hermansky,"Data-driven design of RASTA-like Filters", *Proc.Eurospeech1997*, pp.409-412, Rhodes, Greece,1997

6. N.Kumar, A.G.Andreou,"Generalization of Linear Discriminant Analysis in the Maximum Likelihood Framework", *Proc. Joint Statistical Meeting*, Chicago, August 1996.