MODELISATION AUTOREGRESSIVE EVOLUTIVE ET RECONNAISSANCE DE LA PAROLE

G.Boulianne, G. Chollet *, et Y. Grenier *
INRS-Télécommunications (Univ. du Québec)
3, Place du Commerce
Verdun, Québec CANADA H3E 1H6

Le signal de parole est caractérisé par une alternance de zones spectralement assez stables, entrecoupées de régions transitoires. Les systèmes de reconnaissance proposés par le passé reposent sur des propriétés de stabilité spectrale et de stationnarité; ils obtiennent des performances mitigées pour les régions de transition. Une représentation de ces régions par modèle AR évolutif, valide sur toute la durée d'une région transitoire, est proposée. Les coefficients du modèle dépendent du temps et s'expriment sur une base limitée de fonctions temporelles. Cette méthode de représentation est appliquée à la reconnaissance de segments transitoires C-V extraits de parole naturelle, et comparée à des méthodes plus classiques.

INTRODUCTION

La modélisation autorégressive est bien connue en traitement de la parole sous le nom de prédiction linéaire [1]. Elle oblige à un compromis entre précision et stationnarité qui consiste à découper le signal en fenêtres d'une dizaine de millisecondes.

La modélisation AR évolutive, telle que mise au point par Y.Grenier [2], n'exige pas la stationnarité du signal et de ce fait est mieux adaptée aux régions transitoires de la parole. Le développement d'un espace de représentation adéquat et d'une métrique adaptés à la représentation évolutive fait l'objet de ce travail.

MODELISATION AR EVOLUTIVE

Le modèle AR d'ordre p s'écrit habituellement:

$$y_t + a_1 y_{t-1} + ... + a_p y_{t-p} = b_0 \epsilon_t \qquad (1)$$

Si le processus n'est pas stationnaire, les coefficients a_i deviennent dépendants du temps et sont appelés coefficients évolutifs:

$$y_t + a_1(t-1) y_{t-1} + ... + a_p(t-p) y_{t-p} = b_0(t) \epsilon_t \qquad (2)$$

Leur expansion sur une base de m fonctions du temps s'écrit

$$a_i(t) = \sum_{j=0}^{m-1} a_{ij} f_j(t) \qquad (3)$$

et rend possible leur calcul [2]. Les a_{ij} sont appelés *composants invariants* du modèle évolutif. En représentant les fonctions de la base sous la forme d'un vecteur $F(t) = [f_0(t)\ f_1(t) ... f_{m-1}(t)]$, le modèle évolutif $M(t)$ est obtenu par

$$M^T(t) = A F^T(t) \qquad (4)$$

où la matrice A est formée des composants a_{ij}. La stationnarité n'étant plus nécessaire, un modèle évolutif peut être calculé pour un segment de parole arbitrairement long.

* E.N.S.T., Paris, France

COEFFICIENTS EVOLUTIFS DU CEPSTRE

Pour un modèle stationnaire, les coefficients cepstraux se déduisent des coefficients de prédiction grâce à une relation récursive ([1]). Les coefficients cepstraux évolutifs sont définis par une extension de cette relation:

$$c_i(t) = a_i(t) + \sum_{k=1}^{i-1} \frac{k-i}{i} c_{i-k}(t) a_k(t) \qquad (5)$$

L'expansion des $c_i(t)$ sur une base de m fonctions orthogonales sur l'intervalle τ permet de dériver une approximation des *composants cepstraux invariants*:

$$c_{iq} = a_{iq} + \sum_{k=1}^{i-1} \frac{k-i}{i} \sum_{r=0}^{m-1} c_{(i-k)r} \sum_{s=0}^{m-1} a_{ks} f_{rsq} \qquad (6)$$

pour $i = 1, ..., p$ et $q = 0, ..., m-1$, et où les constantes f_{rsq} peuvent être précalculées:

$$f_{rsq} = \frac{\int_\tau f_r(t) f_s(t) f_q(t) dt}{\int_\tau f_q^2(t) dt} \qquad (7)$$

Le filtre de prédiction doit être stable afin de garantir un comportement raisonnable des composants cepstraux. La stabilisation d'un modèle est effectuée selon la technique de [3], qui consiste à évaluer le polynôme $A(z)$ sur un cercle de rayon supérieur à 1.

DISTANCE ENTRE MODELES EVOLUTIFS

Des essais préliminaires sur la distance euclidienne entre spectres logarithmiques, coefficients de prédiction,de réflexion, et du cepstre ont montré les mêmes tendances pour le modèle évolutif que celles observées par [1] et [4]. La métrique euclidienne sur les coefficients cepstraux a été retenue pour les tests de reconnaissance.

La distance euclidienne entre deux segments de parole décrits par deux trajectoires de paramètres, i.e. deux suites A et B de N points à p dimensions s'écrit habituellement:

$$d(A,B) = \sum_{n=0}^{N-1} \sum_{i=1}^{p} (b_i(n) - a_i(n))^2 \qquad (8)$$

L'équivalent pour deux trajectoires évolutives décrites par des composants invariants est

$$d_e(A_e, B_e) = \sum_{q=0}^{m-1} h_q \sum_{i=1}^{p} (b_{iq} - a_{iq})^2 \qquad (9)$$

où les coefficients h_q dépendent de la base de m fonctions.

$$h_q = \int_\tau f_q^2(t) dt \qquad (10)$$

ANAMORPHOSE TEMPORELLE

Les segments sont modélisés sur l'intervalle τ, subissant une normalisation linéaire du temps. Des déformations non-linéaires peuvent être obtenues directement dans le domaine des composants invariants. Soit la transformation temporelle $t' = u(t)$. Si Γ est une matrice de transformation dont les éléments sont

$$\gamma_{ij} = \frac{\int_\tau f_i(u(t)) f_j(t) dt}{\int_\tau f_j^2(t) dt} \qquad (11)$$

un modèle transformé s'exprimera en fonction de la base originale et de composants transformés $A' = A\Gamma$:

$$M^T(u(t)) = AF^T(t') = A\Gamma F^T(t) = A'F^T(t) \qquad (12)$$

En paramétrisant $u(t)$ par un polynôme $a_0+a_1t+a_2t^2+ ... + a_dt^d$, la matrice devient:

$$\Gamma = \{\gamma_{ij}\} = \{\gamma_{ij}(a_0, a_1, ...a_d)\} \qquad (13)$$

La transformation optimale est celle qui minimise la distance entre un modèle B et un modèle A anamorphosé:

$$d_e^* = \min d_e(A\Gamma, B) = \min\ \Delta(A, B, a_0, ..., a_d) \qquad (14)$$

avec des contraintes qui restreignent aux transformations plausibles: positivité de la pente (pas d'inversion du temps), degré peu élevé du polynôme $u(t)$, intervalle transformé situé dans l'intervalle τ. Le problème se résoud par un algorithme d'optimisation non-linéaire classique.

SEGMENTATION

L'évaluation des techniques précédentes est faite sur des segments transitoires. Leurs frontières ont été définies comme étant les points de pente maximum d'une fonction de stabilité:

$$\nu(n) = \frac{-1}{4} \sum_{i=1}^{3} \sum_{j=-2}^{j=+2} |r_i(n) - r_i(n+j)| \qquad (15)$$

Les $r_i(n)$ sont les coefficients de réflexion d'une fenêtre n. Les régions considérées non-stationnaires sont ainsi celles où la dérivée seconde de la stabilité est positive. Cette définition ne comporte pas de seuils arbitraires ou dépendants du signal.

EXPERIENCES ET RESULTATS

Les expériences ont porté sur des transitions consonnes-voyelle de langue française. Dix séries des 18 syllabes /le/ /ʀe/ /je/ /we/ /ye/ /ve/ /fe/ /se/ /ʃe/ /ʒe/ /me/ /ne/ /pe/ /te/ /ke/ /be/ /de/ /ge/ prononcées par un seul locuteur adulte mâle, en ordre aléatoire, ont été filtrées, numérisées à 12 bits, puis segmentées en régions instables. Parmi ces régions, 175 situées immédiatement avant le noyau vocalique stable ont été extraites et modélisées avec 16 pôles et 4 fonctions de base (polynômes de Legendre). Ces dimensions sont basées sur l'optimisation du critère d'Akaike. Les composants invariants de prédiction ont ensuite été transformés en composants invariants cepstraux. Environ 19% des modèles ont dû être stabilisés. Les modèles cepstraux évolutifs obtenus ont été soumis à quatre expériences:

1. Les dix séries ont été divisées en deux moitiés de 5 séries; la distance euclidienne entre chaque segment d'une moitié et tous les autres segments de l'autre moitié a été évaluée, sans anamorphose. Cette procédure a produit 175 tests de reconnaissance.

2. La procédure 1 a été répétée en introduisant une anamorphose de degré $d = 2$ dans l'évaluation de la distance.

3. Chaque série a été comparée à des références obtenues en combinant les segments des 9 autres séries ("leave one out").

4. La procédure 3 a été répétée avec, pour chaque série, des références auxquelles elle avait participé. Ainsi les segments testés avaient servi à l'apprentissage.

Taux de reconnaissance (175 tests)			
Expérience	Rang du premier correct		
no.	1	2	3
1	57%	66%	70%
2	57%	68%	69%
3	58%	69%	78%
4	72%	82%	88%

La comparaison des expériences 1 et 2 montre que le taux de reconnaissance n'est pas modifié sensiblement par l'anamorphose. Seulement 16% des erreurs commises dans l'une ne le sont pas dans l'autre. L'anamorphose ne dégrade pas la capacité de discrimination de la mesure de distance, mais il ne semble pas y avoir d'avantage à l'utiliser pour des segments aussi courts. L'algorithme de création de références, mis en évidence dans l'expérience 3, se révèle efficace.

Les candidats aux erreurs les plus fréquentes se retrouvent parmi les segments qui ont dû subir une stabilisation. 48% des erreurs sur ceux-ci proviennent d'une confusion avec un autre segment stabilisé. On peut s'attendre à une amélioration marquée du taux de reconnaissance si une autre méthode de stabilisation peut être mise au point, qui n'aplatisse pas l'enveloppe spectrale.

Ces résultats peuvent être comparés aux expériences de [5] sur des séries consonnes-voyelle françaises similaires; 12 systèmes de reconnaissance disponibles sur le marché européen avaient alors obtenu un taux de reconnaissance situé entre 40% et 85%.

CONCLUSIONS

Cette étude montre qu'il est possible de développer, pour la modélisation évolutive, des techniques semblables à celles dont on se sert en prédiction linéaire. L'espace peut être muni d'une métrique utilisable pour la reconnaissance et de transformations permettant une anamorphose temporelle. Les résultats obtenus permettent déjà d'identifier les points faibles des techniques développées, sur lesquels devraient s'attarder de futurs travaux.

BIBLIOGRAPHIE

Markel J.D and A.H Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

Grenier,Y., " Time-dependent ARMA Modeling of Nonstationary Signals", *IEEE Trans. Acoust. Speech and Sign. Proc.*, vol. ASSP-31 no.4, aug. 1983, pp. 899-911.

Haskew J.R., Kelly J.M., Kelly R.M. and T.H. McKinney, "Results of a Study of the Linear Prediction Vocoder", *IEEE Trans. Comm.*, vol. COM-21 no. 9,sept. 1973, pp.1008-1014.

Davis S.B. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust. Speech and Sign. Proc.*, vol. ASSP-28 no.4, aug. 1980, pp. 357-366.

Chollet G.F., Astier A.B.P, and M. Rossi, "Evaluating the Performance of Speech Recognizers at the Acoustic - Phonetic Level", *Proc. ICASSP*, Atlanta 1981. pp.758-761.