

MODELISATION PARAMETRIQUE ET CLASSIFICATION AUTOMATIQUE  
DE SIGNAUX DE FORME TRANSITOIRE

C. VILLEMUR\* - F. CASTANIE\* - B. GEORGEL\*\*

\*ENSEEIH/Gapse 2, Rue Camichel 31071 TOULOUSE CEDEX

\*\*EDF/Direction Etudes et Recherches 6, Quai Watier 78400 CHATOU

RESUME

La présente communication décrit une procédure de classification automatique mise en place sur une banque de signaux tests, préalablement expertisés, de type complexe et issus de contrôle non destructif.

L'étape de classification intervient après une phase de détection-segmentation et une phase de modélisation paramétrique (Prony étendue) des transitoires du signal, représentatifs d'événements que l'on désire classifier. Sur les vecteurs de paramètres obtenus — paramètres autorégressifs, coefficients cepstraux, coefficients de réflexion, tous introduits à partir d'une modélisation autorégressive — on mesure les performances, en pourcentage de bien classés, de plusieurs métriques sur des classes définies a priori. Nous mettons en évidence la supériorité de la métrique associée à la matrice d'inertie — de covariance totale (métrique de Mahalanobis) et interclasse — opérant sur les coefficients cepstraux.

I - INTRODUCTION

Les contrôles par courants de Foucault (C.D.F.) de tubes de générateurs de vapeur (G.V.) permettent la détection de défauts pouvant apparaître dans le métal sous diverses formes. Pour lors, le dépouillement et l'analyse des données, en nombre considérable, sont réalisés uniquement par des opérateurs humains, à l'aide de critères purement visuels: reconnaissance de trajectoires dans le plan de phase.

L'objectif est d'automatiser l'ensemble de la chaîne, de la détection des défauts à leur reconnaissance et ce à l'aide d'outils de modélisation paramétrique du signal\*. Détection-segmentation, modélisation-compactage des données (en vue de l'archivage), classification automatique, pourront donc être implantées sur des appareils informatiques opérant sur site et si possible en temps réel.

II - CARACTERISTIQUES DES SIGNAUX ETUDIES

II-1 - Présentation générale

Les signaux C.D.F. traités ici, présentent la particularité de posséder deux composantes en quadrature. Ils résultent du passage d'une sonde différentielle évoluant à vitesse constante — contrôle non destructif (C.N.D.) — parcourue par un courant alternatif créant un champ magnétique dans la structure métallique à examiner, les tubes de G.V. dans notre cas. L'apparition d'une anomalie, d'un défaut dans le métal ou la présence d'un objet métallique à proximité immédiate du tube perturbe les courants induits et par conséquent, occasionne une variation d'impédance de la bobine: on recueille deux signaux temporels X et Y en quadrature. La réponse enregistrée, prend l'aspect de transitoires 'courts', sur l'une ou l'autre des voies ou sur les deux voies en même temps, lors du passage de la sonde en regard d'une perturbation (voir Fig. 1).

\* Ce travail a été accompli dans le cadre d'une convention INPT/EDF et fait partie de la thèse de doctorat de C. VILLEMUR, membre du GRECO/SARTA.

SUMMARY

This paper describes an automatic classification procedure set up on a base of "test" signals, previously valued, complex-form, issued from eddy-current controls. The classification stage follows a detection-segmentation phase and a phase of parametric modelling (extended Prony) of the transient forms appearing in the signals, representatives of events we intend to classify. On the vectors of computed parameters — autoregressive parameters, cepstral coefficients, reflexion coefficients, all being obtained through the autoregressive modelling — we measure the performances, in terms of percentages of well classified, of several metrics on a priori given classes. We highlight the superiority of the metric based on the covariance matrix — total (Mahalanobis distance) and within-classes — operating on the cepstral coefficients.

II-2 - But du contrôle

On cherche à détecter la présence de discontinuités sachant que celles-ci se superposent et peuvent parfois être masquées par des signaux provenant de la structure même du tube ou de son proche environnement. Divers phénomènes rendent en effet difficile un diagnostic exact de la part de l'expert chargé de dépouiller les enregistrements: — existence d'un bruit de fond "coloré" instable d'origine non électrique, dans la bande passante des signaux défauts qui affecte plus particulièrement la voie X — pics de forte amplitude créés par les plaques de support et de fixation — débuts de courbure des tubes...

III - TRAITEMENT PARAMETRIQUE

Les avantages des méthodes paramétriques sont évidents à toutes les étapes: détection en ligne des transitoires du défaut, reconstruction du signal à partir des coefficients du modèle, en général en faible nombre, d'où archivage compact d'un volume considérable de données, classification automatique sur les divers jeux de paramètres obtenus.

III-1 - La segmentation

Sur une zone 'bruit seul', ou supposée telle, nous appliquons une modélisation autorégressive (AR) complexe de type covariances (algorithme de Marple) et nous suivons en ligne les franchissements d'un seuil  $s$  par l'erreur de prédiction linéaire (EPL), résidu du modèle. La procédure de détection des non stationnarités — des défauts en l'occurrence — est simplement basée sur le temps de séjour du module de l'EPL au-dessus de ce seuil. La stratégie bien connue est la suivante [10] [11]: *si un dépassement dure au moins L échantillons consécutifs, il y a présence d'une non-stationnarité. Dans les autres cas, on ne tient pas compte des franchissements du seuil.*



### Résultats expérimentaux

La Fig. III-3 montre les résultats obtenus pour un type de défaut dont les parties réelles et imaginaires figurent Fig. III-1 et Fig. III-2 respectivement. Nous avons pris :  $p$  (ordre du modèle)=6,  $L=4$  et  $s=2 \cdot \sigma_E$  ( $\sigma_E$  : variance de l'EPL) Aucune fausse alarme n'apparaît et la détection est bien assurée. Les modalités d'implantation — longueur de fenêtre initiale, ordre, etc... — ont déjà discutées par ailleurs [10] [11].

#### Remarque

Les simulations sont opérées sur une banque de signaux "test" [7], comportant des défauts dits "étalons" dont les caractéristiques sont parfaitement connues : localisation dans le tube, largeur, profondeur, orientation.

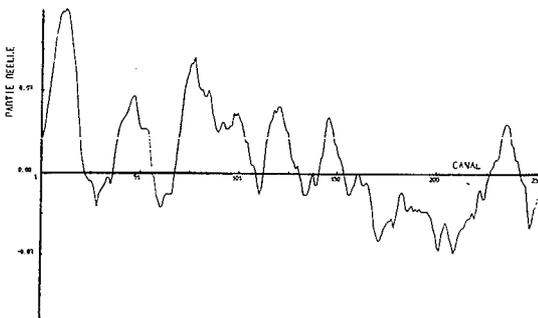


fig III.1

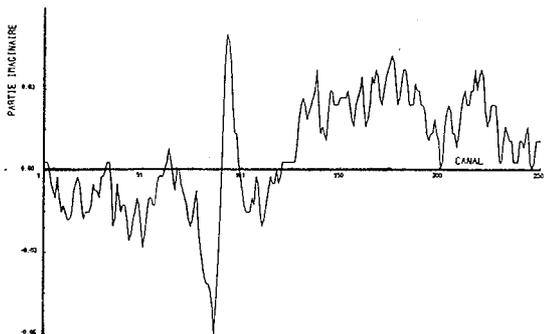


fig III.2

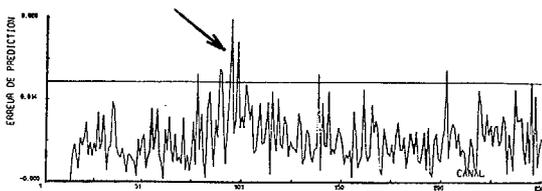


fig III.3

### III-2 La modélisation

Le défaut étant détecté, se pose d'abord le problème de pouvoir le reconnaître parmi les  $K$  classes de défauts déjà répertoriées sur la base test, puis ensuite celui de l'archiver, donc pouvoir le synthétiser à l'aide des coefficients du modèle (ce qui permettra de valider à posteriori l'ordre et le modèle choisis).

— La deuxième étape peut être accomplie en utilisant l'algorithme de Prony étendue qui fait usage dans une première phase du modèle AR précédemment décrit et qui se prête bien à la modélisation des formes transitoires observées. En pratique, on appliquera la méthode de Prony sur une fenêtre courte ( $\sim 128$  points) suivant la détection.

rappel méthode de Prony étendue : soit un signal discret complexe  $x(n)$ , on cherche à l'approximer par  $p$  exponentielles complexes de la forme

$$\hat{x}(n) = \sum_{i=1}^p b_i \cdot z_i^n \quad (3.1)$$

les  $b_i$ ,  $z_i$  contiennent les informations de phase, amplitude et de fréquence, amortissement.

La première étape, le processus de classification nécessite de décrire chaque forme, les transitoires, par un modèle et donc lui associer un vecteur formé de  $p$  paramètres descriptifs, tous introduits à partir du modèle AR, auxquels nous ajouterons éventuellement une  $(p+1)$ ème composante représentative de la puissance moyenne (voir IV-1).

La validation du modèle peut être effectuée à l'aide des outils propres à la classification automatique appliquée à une famille de formes étiquetées (expertisées). C'est l'objet du chapitre suivant.

## IV - CLASSIFICATION

Dans un premier temps, on s'attachera à trouver l'algorithme de classification — choix d'un espace de représentation du signal et d'un processus décisionnel optimal — donnant les meilleurs résultats sur la base test puis on simulerà la reconnaissance d'un défaut "inconnu" n'appartenant pas à cette base (il peut s'avérer nécessaire de pré-traiter le signal: fenêtre de Hamming, pré-accentuation, transformée logarithmique, etc). Une démarche similaire a été adoptée pour la classification dans des domaines aussi divers que l'électromyographie, la parole ou le traitement des fouillis en radar [4] [2] [1].

### IV-1 - Formulation du problème

Soit  $K$  classes  $O_i$  connues dont on possède  $K$  modèles  $M_i$  dans  $\mathbb{C}^{p+1}$  (les vecteurs associés aux centres de gravité): ayant un objet — un défaut —  $O_j$  à reconnaître et à classer dans un ensemble à  $N$  éléments — la base test — comment décider :  $O_j \in O_i$  ?

Une mesure évidente de la similarité de 2 formes et par conséquent de deux défauts, est la distance qui sépare les vecteurs les représentant. Cette distance peut être calculée à l'aide de plusieurs métriques, n'ayant pas forcément une interprétation spectrale.

$$O_i \in O_j \Leftrightarrow d(X_i, Y_j) = \min_{m=1, \dots, K} d(X_i, Y_m) \quad (4.1)$$

$X_i$ : vecteur représentant  $O_i$

$X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,p}, (e)]$

$Y_j$ : centre de gravité de la  $j$ ème classe

$p$ : ordre du modèle AR

On prend  $(e)$  égal à la puissance du signal  $R(e)$  ou bien  $\sigma_e$  ou encore  $\log(\sigma_e)$  dans le cas des coefficients cepstraux. Pour  $d$ , la symétrie et l'inégalité triangulaire ne sont pas requises.

### IV-2 - Détails des métriques utilisées

Elles sont toutes dérivées de la forme:

$$d(X, Y) = (X - Y)^T \cdot P \cdot (X - Y)^* \quad (4.2)$$

\* dénote le complexe conjugué,  $X^T$  le vecteur transposé de  $X$  et  $P$  une matrice de pondération symétrique, définie positive.



$\alpha$  / distance euclidienne:  $P=I_p$  (4.3)

$\beta$  / pondération par les variances:  $P=D, D$  étant une matrice diagonale ( $d_1, \dots, d_p$ )

\* variances totales:  $d_i=1/\sigma_i^2$  en posant

$$\sigma_i^2=1/N \cdot \sum_{k=1}^N (x_{ki} - \bar{x}_i)(x_{ki} - \bar{x}_i)^* \text{ et } \bar{x}_i=1/N \cdot \sum_{k=1}^N x_{ki} \quad (4.4)$$

\* variances interclasses:  $d_i=1/\lambda_i^2$  avec

$$\lambda_i^2=1/K \cdot \sum_{k=1}^K (y_{ki} - \bar{y}_i)(y_{ki} - \bar{y}_i)^* \text{ et } \bar{y}_i=1/K \cdot \sum_{k=1}^K y_{ki} \quad (4.5)$$

\* variances intra-classes:  $d_i=1/\mu_i^2$  avec  $\mu_i^2=\sigma_i^2-\lambda_i^2$  (th. de Huyghens) (4.6)

$\gamma$  / pondération par les matrices de covariance: distance de 'Mahalanobis' [1] [8] ;  $P = T^{-1}$ ,  $T = (t_{ij})$  ( on parle indifféremment de matrice de covariance totale ou de matrice d'inertie du nuage des N points), les  $t_{ij}$  sont estimés de façon biaisée comme suit:

$$t_{ij} = \text{partie réelle} (1/N \cdot \sum_{k=1}^N (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)^*) \quad (4.7)$$

**Remarque**

On met l'accent sur les composantes matricielles où les variances sont les plus élevées, c'est-à-dire là où les dispersions entre classes sont les plus importantes.

On peut également utiliser la matrice des covariances intra-classes W ( de l'anglais within) calculée d'après la formule (4.6):  $W=T-B$ ,  $B = (b_{ij})$

$$b_{ij} = \text{partie réelle} (1/K \cdot \sum_{k=1}^K (y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j)^*) \quad (4.8)$$

$\delta$  / distance cepstrale avec dérivée de phase:

$$P = \text{diag} (\dots, i^2, \dots)$$

$$d(X, Y) = \sum_{i=1}^P i^2 (c_{xi}^2 - c_{yi}^2)^2 \quad (4.9)$$

La distance cepstrale classique — également appelée RMS Log [3] — est en fait une distance euclidienne généralisée sur 2p coefficients cepstraux. C'est celle [2] qui donnait les meilleurs taux de reconnaissance en parole.

**V - RESULTATS DES SIMULATIONS**

Dans un premier temps, nous allons présenter sous la forme d'un tableau à deux entrées — ordre de modélisation / dimension de l'espace de représentation et métrique utilisée — les résultats obtenus en pourcentages de bien classés sur la banque de signaux test, que nous avons subdiviser en 12 classes. Certaines d'entre elles résultent de l'éclatement d'une classe 'mère' en plusieurs sous-classes. Il apparait intuitif que ce sont les classes possédant le plus de variance qui viendront altérer le plus le classement des signaux. Pour chaque classe le vecteur moyen est calculé à partir d'au moins trois éléments. Au total près de 49 signaux complexes font partie du nuage.

Ces simulations nous fourniront la combinaison — prétraitement ( voir remarque ci-dessous) + paramètres + ordre + métrique — optimale. Il nous restera par la suite à simuler le processus de reconnaissance sur des signaux 'inconnus', n'appartenant pas à la base test et donc ne contribuant pas aux calculs des variances. Ces signaux ont été bien sûr préalablement expertisés, ceci nous permettra de pouvoir conclure quant à l'exactitude du diagnostic en sortie de l'algorithme.

**Remarque**

Dans chaque case des tableaux, figurent sur la 1ère ligne les % de bien-classés obtenus avec les paramètres AR ( $a_k$ ), sur la 2ième ceux obtenus avec

les coefficients de réflexion ( $k_i$ ) et enfin sur la dernière ceux fournis par les coefficients cepstraux ( $c_k$ ).

Tableau 1: p paramètres uniquement.

Tableau 2: p paramètres + 1 composante énergétique moyenne ( $\sigma_{E^2}$  ou  $\text{Log}(\sigma_{E^2})$ ).

( nous avons remarqué que, par rapport à une fenêtre rectangulaire classique, une fenêtre de Hamming améliorerait les % sur les  $c_k$  dans tous les cas alors qu'elle les dégradait sur les  $a_k$  et les  $k_i$ , en conséquence nous avons conservé cette fenêtre mieux adaptée au jeu des paramètres optimaux — les  $\{c_k\}$ , à vérifier a posteriori — ).

	$D_E$	$D_\sigma$	$D_\mu$	$D_T$	$D_W$	$D_{rms}$	$D_{der}$
p= 5	49	61	54	72	74	--	--
	54	63	57	72	74	--	--
	62	67	64	70	72	64	74
p=10	64	62	59	83	87	--	--
	64	66	62	83	87	--	--
	64	79	79	83	90	77	81
p=14	67	69	72	85	87	--	--
	67	72	72	85	87	--	--
	69	87	87	87	90	79	88
p=18	74	74	77	88	92	--	--
	74	77	81	88	90	--	--
	74	87	92	92	94	77	90

	$D_E$	$D_\sigma$	$D_\mu$	$D_T$	$D_W$	$D_{rms}$	$D_{der}$
p= 5	46	59	62	75	79	--	--
	54	66	66	75	79	--	--
	74	77	72	77	82	64	74
p=10	67	64	64	87	90	--	--
	67	62	72	81	85	--	--
	72	87	87	89	92	77	81
p=14	67	77	79	88	92	--	--
	64	77	81	88	90	--	--
	77	88	88	94	97	79	88
p=18	74	77	77	94	94	--	--
	74	77	81	94	94	--	--
	74	90	90	97	100	77	90

$D_E$ : dist. euclidienne;  $D_\sigma$ : pond. par les variances totales;  $D_\mu$ : pond. par les variances intra-classes;  $D_T$ : pond. par la matrice  $T^{-1}$ ; idem pour  $D_W$  avec  $W^{-1}$ ;  $D_{rms}$ : dist cepstrale classique;  $D_{der}$ : dist. cepstrale avec dérivée de phase.

Au vu des résultats obtenus, nous avons donc choisi pour la suite la métrique de 'Mahalanobis' appliquée à p+1 coefficients cepstraux avec un ordre optimum se situant entre 15 et 20. En ce qui concerne la nature des paramètres les plus pertinents pour une classification automatique, ceci vient corroborer les résultats déjà acquis en traitement de la parole [2] et en cours d'évaluation pour l'E.M.G. [4].

Sur une trentaine de signaux 'inconnus', le pourcentage de reconnaissance correcte 'n'est' que de 83%. Toute tentative en vue d'augmenter l'ordre se solde par des instabilités numériques: matrices singulières, mauvaise modélisation, etc. Toutefois nous notons que, lorsqu'il y a confusion entre deux classes, ceci se produit en général lorsque le bruit est très important ou bien quand les allures temporelles respectives sont très proches. Les centres de gravité de certaines classes sont si proches, que dans le cas où il n'y a pas eu reconnaissance, si l'on décide de l'appartenance à une classe, non pas à l'aide de la distance minimale (4.1) mais à l'aide de la distance immédiate supérieure, on observe alors une reconnaissance quasi-parfaite (94%).

Il ne faut pas perdre de vue que nous nous proposons essentiellement pour l'instant de fournir une aide au diagnostic.



## VI - CONCLUSION

Le traitement paramétrique du signal s'avère donc être un outil particulièrement puissant pour la reconnaissance des formes.

De plus, les résultats encourageants enregistrés laissent augurer de la possibilité de tester cet algorithme en temps réel. Le principal coût calculatoire étant dû à l'obtention de la matrice des covariances et de son inverse, cette dernière doit être acquise au préalable — on peut concevoir de l'actualiser au fur et à mesure de la reconnaissance d'un objet — : le coût algorithmique se réduit alors à celui de la modélisation AR (proportionnel en général à  $Mxp^2$ ,  $M$  étant le nombre de points du signal) et à celui du calcul de la distance ( $Kxp^2$  additions et autant de multiplications pour 'Mahalanobis').

### mots-clés:

traitement paramétrique du signal, signaux C.D.F., classification automatique, métrique de 'Mahalanobis'.

[12] S.KAY - L.MARPLE: 'Spectrum analysis - A modern perspective' dans Proc. of the IEEE, vol 69, n°11, Nov 1981, pp 1380-1418.

[13] L.MARPLE: 'A new autoregressive spectrum analysis algorithm' dans IEEE Trans. on ASSP, vol 28, n°4, Août 1980, pp 441-454

[14] Y.GRENIER-L.MICLET-KAZILEWSKI: 'Reconnaissance de formes', cours E.N.S.T. Paris 1978.

## Bibliographie

[1] J.F. AGNEL: 'Elimination de fouillis en radar et détection séquentielle de cibles: une approche par la reconnaissance des formes et le filtrage en treillis' -thèse de docteur ingénieur-E.N.S.T. Paris-Fév 1985.

[2] Y.GRENIER: 'Modélisation et reconnaissance de la parole' dans 'Outils et modèles mathématiques pour l'automatique, l'analyse et le traitement du signal', Editions du C.N.R.S., T 2, pp 617-637, 1982.

[3] A.H. GRAY, jr - J.D. MARKEL: 'Distance measures for speech processing' dans IEEE TRANS ON ASSP, vol 24, n°5, Oct 1976, pp 380-391.

[4] S.SITBON et al.: 'Classification et aide au diagnostic en E.M.G.', GAPSE/ENSEEIH, article à paraître GRETSI 87.

[5] P.V. de SOUZA: 'Statistical tests and distance measures for LPC coefficient' dans IEEE TRANS ON ASSP, vol 25, n°6, Dec 1977, pp 554-559.

[6] X.L.WANG: 'Paramétrisation, classification et reconnaissance de formes', rapport de D.E.A GAPSE/ENSEEIH, Juin 1986.

[7] A.ROUSSILLO-B.GEORGEL: 'Une ébauche de banque de signaux pour la reconnaissance automatique de défauts par courants de Foucault', Note technique externe, EDF/Chatou, Juin 1986.

[8] DIDAY-LEMAIRE-POUGET-TESTU: 'Eléments d'Analyse de Données', pp 309-334, éd. DUNOD.

[9] C.VILLEMUR: rapports de contrat GAPSE/ENSEEIH-EDF Etudes et Recherches, 1985-87.

[10] P.SOULE - F.CASTANIE: 'Méthode de détection de non-stationnarités par analyse des propriétés de franchissement d'un seuil par l'erreur de prédiction linéaire', GRETSI 1983, pp 899-903.

[11] E.DAYMIER -F. CASTANIE: 'Analyse d'une méthode de détection des sauts de moyenne de variance', GAPSE/ENSEEIH, GRETSI 1985.