Coopération entre codeurs neuro-prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes

Cooperation between neural predictive coders for feature extraction in phoneme recognition

M. Chetouani, B. Gas, J.L. Zarader
Laboratoire des Instruments et Systèmes d'Ile-De-France
Université Paris VI
BP 164, Tour 22-12 2ème étage
4 Place Jussieu, 75252 Paris Cedex 05, France
mohamed.chetouani@lis.jussieu.fr gas@ccr.jussieu.fr zarader@ccr.jussieu.fr

Résumé

L'extraction de caractéristiques est une des plus importante étape dans le processus de reconnaissance de la parole. Dans ce papier, nous présentons une architecture neuronale appelée Cooperative Modular Neural Predictive Coding (CMNPC). Elle est basée sur l'interaction de codeurs discriminants DFE-NPC (Discriminant Feature Extraction) optimisés pour le traitement d'un groupe phonétique (macro-classe) par le biais d'un critère : le Rapport d'Erreur de Modélisation (MER). Nous proposons une validation théorique par une description statistique de l'architecture. Les performances de l'architecture sont estimées sur une tâche de reconnaissance de phonèmes. Ces phonèmes sont extraits de la base Darpa-Timit. Nous comparons les performances avec les méthodes de codage traditionnelles : LPC, MFCC et PLP. Nous proposons également une étude expérimentale de l'influence de la méthode de coopération sur les taux de reconnaissance.

Mots Clef

Extraction de caractéristiques, Reconnaissance de phonèmes, Coopération entre les architectures neuronales.

Abstract

Speech feature extraction is one of the most important stage in the speech recognition process. In this paper, we propose a new neural networks architecture called the Cooperative Modular Neural Predictive Coding (CMNPC). It is based on the interaction of discriminant experts DFE-NPC (Discriminant Feature Extraction) optimized for the processing of one phonetic group (macro-class) by the help of a criterion : the Modelisation Error Ratio (MER). We propose a theoretical validation by a statistic description of the architecture. The performances of this architecture are estimated in a phoneme recognition task. The phonemes are extracted from the Darpa-Timit speech database. Comparisons with traditional coding methods (LPC, MFCC, PLP) are presented. They put in obviousness an improvement of the recognition rates. We also propose an empirical study of the influence of the cooperation method on recognition rates.

Keywords

Feature extraction, Phoneme recognition, Cooperation between neural architectures.

1 Introduction

En reconnaissance de la parole, l'extraction de caractéristiques peut-être réalisée de plusieurs manières. En effet, les vecteurs codes sont couramment extraits à l'aide de méthodes temporelles comme le codage linéaire prédictif (Linear Predictive Coding LPC) ou de méthodes cepstrales comme le codage MFCC (Mel Frequency Cepstral Coding). Le codage PLP (Perceptual Linear Predictive coding) est un exemple de l'application des connaissances du système auditif humain en reconnaissance de la parole.

L'extraction de caractéristiques est un élément clé pour la mise au point d'un système de reconnaissance. De nombreux travaux ont montré l'importance de cette étape [4, 11, 13]. Les nouvelles méthodes proposées sont basées soit sur l'analyse de données comme l'Analyse en Composantes Indépendantes (ACI) [14], mais aussi sur l'analyse multi-résolution [8]. L'ACI offre la possibilité d'extraire des moments statistiques d'ordre supérieures (Higher-Order Statistics HOS). Le principal avantage de l'analyse multi-résolution comme de la transformée en ondelettes est la résolution temps-fréquence qui permet une meilleure localisation des caractéristiques.

Notre approche est une extension au domaine non linéaire du codage LPC par les réseaux de neurones : le Codage Neuro-Prédictif (Neural Predictive Coding NPC) [7]. Cette méthode est en adéquation avec le fait que le processus de

production de la parole est non linéaire [19].
Le travail proposé est une application du principe "diviser pour mieux régner" à l'extraction de caractéristiques. Traditionnellement, l'extraction de caractéristiques est réalisée de la même manière pour les différents types de phonèmes sans considérer leurs différences. Une méthode pour améliorer l'extraction de caractéristiques peut-être d'adapter le traitement à chaque type de phonème. Cette stratégie a été utilisée pour l'estimation de probabilités à posteriori de phonèmes [18].
Dans ce papier, nous proposons une nouvelle approche d'apprentissage pour la coopération d'experts discriminants. Un des problèmes avec l'apprentissage de comité d'experts est le manque de discrimination entre les experts. Chacun des modules a une "perception" différente de l'espace d'entrée [2]. Une des méthodes proposées est la corrélation négative des experts : Negative Correlation Learning (NCL) [15]. Le principe de cette méthode d'apprentissage consiste à introduire un terme de pénalité dans la fonction d'erreur de chacun des réseaux. Ce terme de pénalité est estimé à partir de la corrélation entre les sorties des réseaux. Cette méthode permet d'entraîner de manière simultanée les réseaux.
La méthode que nous proposons est différente. Elle consiste à ajouter une classe à chacun des réseaux. Cette classe représente les "autres groupes" phonétiques. Chaque codeur NPC expert est optimisé sur un groupe phonétique (macro-classe) avec une "perception" des autres experts.
Dans un premier temps, nous présentons le codage NPC pour l'extraction de caractéristiques discriminantes. Ensuite, l'architecture coopérative : The Cooperative Modular NPC architecture (CMNPC) est décrite. Nous présentons également une validation théorique du modèle. Les sections suivantes présentent les conditions expérimentales ainsi que les résultats en classification de phonèmes. Finalement, nous donnerons quelques conclusions sur le travail proposé.

2 Codage Neuro-Prédictif

Le codage neuro-prédictif (Neural Predictive Coding) est une extension non linéaire du codage LPC. Le modèle NPC est basé sur un perceptron à une couche cachée utilisé en prédicteur non linéaire (cf. Fig. 1).

2.1 Prédiction non linéaire

De nombreux travaux ont montré la nécessité des prédicteurs non linéaires en traitement de la parole [9, 20]. Les méthodes de prédiction non linéaires peuvent être décomposées en deux types : les méthodes non paramétriques et les méthodes paramétriques. Les méthodes non paramétriques ne nécessitent pas la définition de la fonction de prédiction. Une des méthodes consiste à pré-calculer une table numérique (lookup table) de la fonction afin d'effectuer une prédiction dite "codebook prediction" [21]. L'intérêt principal de cette méthode est la réduction des temps de calcul. Les méthodes paramétriques sont essentiellement

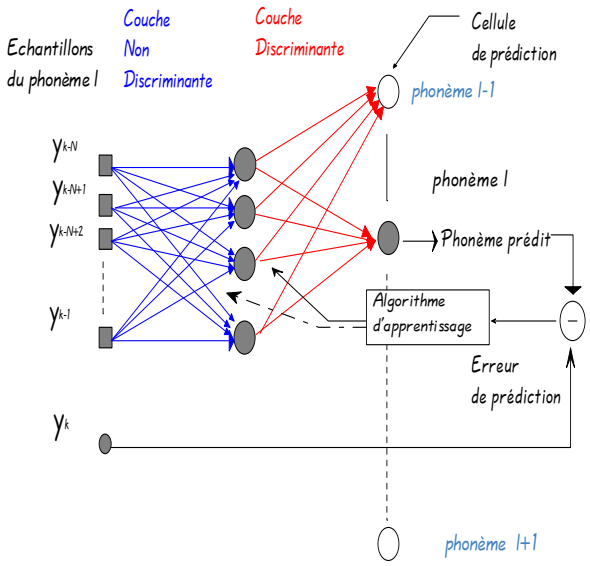

FIG. 1 – Architecture du codeur NPC.

basées sur les filtres de Volterra [20] et les réseaux de neurones [9]. L'avantage principal des filtres de Volterra est que, comme dans le cas des prédicteurs linéaires, les coefficients solutions de l'erreur quadratique minimale peuvent être exprimés de manière analytique. Par contre, le nombre de coefficients augmente rapidement avec la dimension de la fenêtre de prédiction. Cet inconvénient se retrouve également dans le cas des réseaux de neurones, et de plus les poids solutions ne peuvent être exprimés de manière analytique en fonction de l'erreur quadratique. Les réseaux de neurones ont été appliqués avec succès à la prédiction non linéaire de signaux de parole : MLP [9], réseaux à base radiale (RBF) [3] et réseaux récurrents [22]. Le principal avantage d'une approche connexioniste est le fait que les types de non linéarités (l'ordre : 2ème, 3ème, etc. ...) à modéliser ne sont pas fixés [9]. En effet, dans le cas des filtres de Volterra, les non linéarités modélisées sont de type quadratique, et un ordre supérieur ne ferait qu'augmenter le nombre de coefficients.

2.2 Le modèle NPC

Le modèle NPC apporte une solution au problème de dimension du vecteur code des prédicteurs non linéaires tout en gardant l'avantage des réseaux de neurones. La solution consiste à ne considérer comme vecteur code uniquement les poids de la seconde couche.
La fonction de prédiction du modèle NPC est la suivante :

$$\hat{y}_k = F(\mathbf{y}_k) \tag{1}$$

Où k est l'indice des échantillons du signal de parole, $\mathbf{y}_k$ est le contexte de prédiction : $\mathbf{y}_k = [y_{k-1}, y_{k-2}, ..., y_{k-\lambda}]^T$. λ est la dimension de la fenêtre de prédiction .

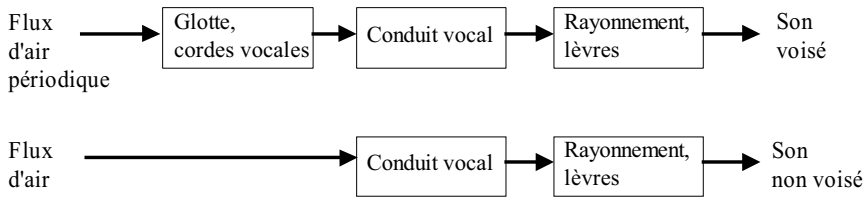

FIG. 2 – Processus de production de la parole dans le cas des phonèmes voisés et non voisés.

F est une fonction non linéaire composée par deux fonctions $G_{\mathbf{w}}$ ($\mathbf{w}$ poids de la première couche) et $H_{\mathbf{a}}$ ($\mathbf{a}$ poids de la couche de sortie) :

$$F_{\mathbf{w},\mathbf{a}}(\mathbf{y}_k) = H_{\mathbf{a}} \circ G_{\mathbf{w}}(\mathbf{y}_k) \tag{2}$$

Avec $\hat{y}_k = H_{\mathbf{a}}(\mathbf{z}_k)$ et $\mathbf{z}_k = G_{\mathbf{w}}(\mathbf{y}_k)$.
Cette méthode de codage nécessite le pré-calcul des poids de la première couche $\mathbf{w}$ et donc l'estimation de la fonction $G_{\mathbf{w}}$. Ceci est fait durant la phase de paramétrisation. Un autre élément important du modèle NPC est l'intégration de la notion de classes lors de cette phase. En effet, la première couche est commune à tous les trames (fenêtre d'analyse) de l'ensemble des phonèmes tandis que la seconde couche est dédiée aux trames d'une seule classe de phonème :

$$Q = \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{l=1}^{M} \left(y_{i,k} - F_{\mathbf{w},\mathbf{a}_l}(\mathbf{y}_{i,k})\right)^2 \delta_{C_i - l} \tag{3}$$

Le symbole de Kronecker associe la classe C_i à la couche de sortie l.
A la suite de la phase de paramétrisation, la première couche du codeur est fixée. Il faut maintenant déterminer les poids de la seconde couche $\mathbf{a}$. Les poids de la seconde couche forment le vecteur code du phonème.

2.3 Principe d'extraction de caractéristiques

Le processus de production de la parole [16, 17] (cf. figure 2) est différent selon le type de phonème (voisé, non voisé, etc. ...). Dans le cas des phonèmes voisés le flux d'air p est un train d'impulsion de période N. Ce flux d'air est modifié par les contributions glottales g, rayonnement (lèvres) r et celle du conduit vocal v. Le signal de parole y résultant est la convolution de p par les réponses impulsionnelles g, r, v des trois parties du processus de production de la parole :

$$y = p * g * v * r \tag{4}$$

Dans le cas de la production de sons non voisés, les cordes vocales ne vibrent pas. Le flux d'air u est considéré comme un bruit blanc :

$$y = u * v * r \tag{5}$$

A ces différents modèles (4 et 5), il faut associer d'autres modèles de production comme dans le cas des nasales. En effet, il est également nécessaire de modéliser la contribution de la voie nasale.
Le modèle de production de la parole peut donc être décomposé en deux parties : une partie commune (le conduit vocal), et des zones spécifiques (contributions glottales et rayonnement des lèvres). Pour des phonèmes ayant des caractéristiques proches, c'est-à-dire pour lesquels le processus de production est pratiquement identique, l'extraction de caractéristiques optimales consisterait à extraire seulement les caractéristiques discriminantes.
Le modèle NPC intègre ce concept : les poids de la première couche $\mathbf{w}$ sont communs à tous les phonèmes tandis que les poids de la seconde couche $\mathbf{a}_l$ sont spécifiques à chaque classe de phonèmes.
Considérons i et j deux allophones [1] appartenant respectivement aux classes de phonèmes différentes C_i et C_j. Les modèles NPC associés à ces deux allophones sont les suivants :

$$\begin{cases} F_{\mathbf{w},\mathbf{a}_i} = H_{\mathbf{a}_i} \circ G_{\mathbf{w}} \\ F_{\mathbf{w},\mathbf{a}_j} = H_{\mathbf{a}_j} \circ G_{\mathbf{w}} \end{cases} \tag{6}$$

Les modèles NPC $F_{\mathbf{w},\mathbf{a}_i}$ et $F_{\mathbf{w},\mathbf{a}_j}$ sont différents tandis que $G_{\mathbf{w}}$ est commune aux deux allophones i et j. Cette fonction $G_{\mathbf{w}}$ rassemble donc les caractéristiques communes alors que les caractéristiques discriminantes sont représentées par les fonctions $H_{\mathbf{a}_i}$ et $H_{\mathbf{a}_j}$. En généralisant à tous les allophones de tous les phonèmes, et sous les hypothèses d'apprentissage du modèle NPC, on peut considérer que les poids de la première couche $\mathbf{w}$ modélisent les caractéristiques communes des classes de phonèmes tandis que les poids des secondes couches $\mathbf{a}$ modélisent les caractéristiques discriminantes.

3 Rapport d'Erreur de Modélisation

Le Rapport d'Erreur de Modélisation (Modelisation Error Ratio MER) [7] est un critère qui permet d'introduire de la discrimination explicite entre les modèles NPC. En effet, les méthodes prédictives présentent un manque de discrimination. Ceci est dû au fait que les modèles sont entraînés de manière indépendante, par conséquent il n'y a pas de discrimination explicite. Le modèle DFE-NPC (Discriminant Feature Extraction) est une extension du modèle NPC intégrant le MER. Le MER est définie comme un rapport d'erreurs de prédiction :

[1] Réalisation acoustique d'un phonème

Lieu	Mode d'articulation						
	Semi-voyelles	Liquides	Nasales	Occlusives (voisés)	Occlusives (non voisés)	Fricatives (voisés)	Fricatives (non voisés)
Antérieure *Bilabiales* *Labiodentales*	w		m	b	p	v	f
Centrale *Alvéolaires* *Palatales*	y	l r	n	d	t	z jh	s ch
Postérieure *Vélaires*			ng	g	k		

TAB. 1 – Lieu et mode d'articulation des consonnes.

Soit L_j^i l'erreur de prédiction du phonème i issue du modèle NPC $H_{\mathbf{a}_j}$ qui est associé au phonème j :

$$L_j^i = \sum_k (y_{i,k} - H_{\mathbf{a}_j} \circ G_{\mathbf{w}}(\mathbf{y}_{i,k}))^2 \quad (7)$$

Le i-MER est le rapport inverse de l'erreur de prédiction du phonème i par le modèle NPC correct et des erreurs de prédiction du même phonème i par les autres modèles NPC :

$$\Gamma_i = \frac{\sum_{j=1,j\neq i}^{M} L_j^i}{L_i^i} \quad (8)$$

Où M est le nombre de classes.
La maximisation du i-MER est utilisée pour la discrimination entre des modèles NPC. En accord avec la définition du i-MER (8), le MER est une extension aux M classes :

$$\Gamma = \frac{Q^d}{(M-1)Q^m} \quad (9)$$

Où $Q^d = \sum_{i=1}^{M} \sum_{j=1,j\neq i}^{M} L_j^i$ et $Q^m = \sum_{i=1}^{M} L_i^i$.
Le modèle DFE-NPC [7] est basé sur la minimisation de l'inverse du MER :

$$Q_{\text{DFE-NPC}} = \frac{1}{\Gamma} \quad (10)$$

Le modèle DFE-NPC a été appliqué avec succès à l'extraction de caractéristiques [7]. Dans ce papier, nous proposons l'étude d'une nouvelle architecture pour la combinaison de modèles DFE-NPC. Comme nous l'avons vu, le principe d'extraction de caractéristiques de ce modèle est de localiser les caractéristiques communes sur les poids de la première couche tandis que les caractéristiques discriminantes sont localisées sur les poids de la seconde couche. Le problème qui se présente lorsque l'on traite un grand nombre de phonèmes est la capacité du réseau à modéliser les caractéristiques communes. La modularité est une des solutions pour résoudre ce problème car elle permet de combiner des experts pour chaque type de phonème.

4 Architecture coopérative

4.1 Description

La nouvelle architecture est basée sur la coopération de modèles experts DFE-NPC. L'interaction est introduite par un MER coopératif. L'idée principale est d'appliquer le principe "diviser pour mieux régner" à l'extraction de caractéristiques. Il s'agit donc de décomposer la tâche d'extraction de caractéristiques. La décomposition consiste à regrouper des classes dans une même macro-classe qui sera traitée par un expert.
Dans le cas de l'extraction de caractéristiques, nous pouvons utiliser l'expertise phonétique. En effet, l'Alphabet Phonétique International fournit une décomposition en groupes phonétiques. Cependant dans une tâche de discrimination, il est intéressant de choisir des groupes ayant des discriminations intrinsèques. Le tableau 1 présente une classification des consonnes en fonction de leurs lieux et modes d'articulation (toutes les consonnes ne sont pas représentées). Les groupes phonétiques peuvent être définis en fonction du lieu ou du mode d'articulation. Le choix est très important car il permet de concentrer la discrimination sur certaines ambiguïtés. Par exemple dans le cas des occlusives, il vaut mieux choisir une partition en fonction du mode d'articulation car la principale différence est le voisement (caractéristique beaucoup plus discriminante). Le lieu d'articulation n'est pas une caractéristique appropriée pour discriminer les consonnes. En effet, le mode d'articulation est une caractéristique beaucoup plus discriminante car il existe de grandes différences entre les consonnes voisées, non voisées, orales, nasales, etc. ...
Par contre, dans le cas des voyelles (cf. tableau 2), le lieu d'articulation est une caractéristique discriminante. En effet, l'aperture de la bouche est une caractéristique difficile à extraire comparativement au lieu d'articulation.

Aperture	Lieu d'articulation		
	Antérieure	Centrale	Postérieure
Fermée	ih		uw
Semi-fermée	ey	ah	ow
Semi-ouverte	eh	er	uh
Ouverte	ae		aa

TAB. 2 – Aperture et lieu d'articulation des voyelles.

L'expertise fournie par la phonétique permet de définir des macro-classes (groupes phonétiques) décrites tableau

3. Les diphtongues qui sont un cas particulier des voyelles sont également traitées.
La décomposition permet d'exploiter ainsi des discriminations intrinsèques entre les macro-classes. Ensuite, chacun des experts aura la tâche d'extraire des caractéristiques discriminantes au sein de sa propre macro-classe. Ceci aura pour effet de fusionner des discriminations issues de différents niveaux. Par exemple, dans le cas des consonnes, les macro-classes sont définies selon leurs modes d'articulation. L'expert d'une de ces macro-classes devra extraire des caractéristiques discriminantes selon le lieu d'articulation.

Ω_1	Voyelles antérieures : ih ey eh ae
Ω_2	Voyelles centrales : ah er
Ω_3	Voyelles postérieures : uw uh ow aa
Ω_4	Diphtongues : ay aw oy
Ω_5	Semi-Voyelles : y w
Ω_6	Liquides : l r
Ω_7	Nasales : m n ng
Ω_8	Occlusives (voisées) : b d g
Ω_9	Occlusives (non voisées) : p t k
Ω_{10}	Fricatives (voisées) : v z jh
Ω_{11}	Fricatives (non voisées) : f s ch

TAB. 3 – Décomposition de la tâche d'extraction de caractéristiques en macro-classes phonétiques

4.2 Modèle coopératif

L'idée clé est d'entraîner le modèle DFE-NPC optimisé pour la macro-classe Ω_τ avec un terme de pénalité représentant les trames des autres macro-classes. Ces trames constituent une nouvelle classe. Dans le cas de la macro-classe des voyelles antérieures constituée des phonèmes suivants : ih, ey, eh et ae, l'apprentissage avec MER coopératif consiste à discriminer ces phonèmes entre eux mais aussi à les discriminer de tous les autres phonèmes en introduisant une cinquième classe : "les autres macro-classes".
Pour un DFE-NPC expert de la macro-classe Ω_τ, on définit le i-MER Coopératif pour un phonème de la macro-classe Ω_τ :

$$\Psi_i = \frac{\sum_{j=1,j\neq i}^{M} L_j^i + L_{\Omega_{\overline{\tau}}}^i}{L_i^i} \qquad (11)$$

Où $L_{\Omega_{\overline{\tau}}}^i$ est l'erreur de prédiction du phonème i par le modèle $H_{\mathbf{a}_{\Omega_{\overline{\tau}}}}$ associé aux autres macro-classes $\Omega_{\overline{\tau}}$.
Comme dans le cas du MER (9), le MER Coopératif Ψ est une extension du i-MER Coopératif à l'ensemble des M classes. L'optimisation du nouveau modèle DFE-NPC Coopératif est basée sur la minimisation de l'inverse du MER Coopératif Ψ. En effet, comme les trames des macro-classes $\Omega_{\overline{\tau}}$ peuvent-être considérées comme une classe ajoutée, l'apprentissage d'un modèle DFE-NPC coopératif sur M classes revient à un apprentissage d'un modèle DFE-NPC sur $N = M + 1$ classes :

$$Q^M_{\text{Cooperative DFE-NPC}} \Longleftrightarrow Q^N_{\text{DFE-NPC}} \qquad (12)$$

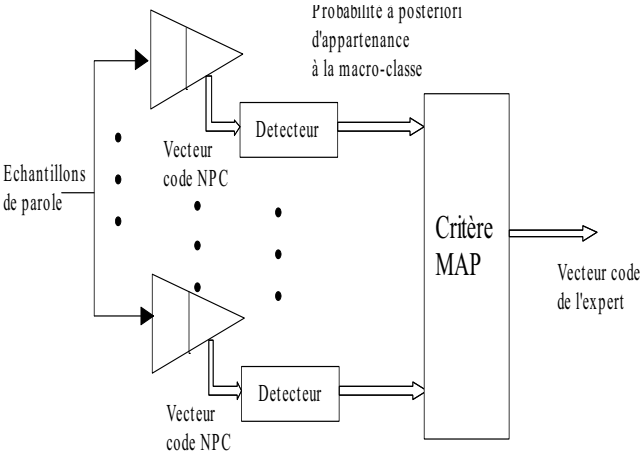

FIG. 3 – Architecture CMNPC : DFE-NPC experts et détecteurs de macro-classes.

4.3 L'étape de macro-classification

Le principe de macro-classification est basé sur l'association à chaque expert d'un détecteur. L'expert DFE-NPC associé à la macro-classe Ω_τ encode le signal de parole et fournit un vecteur caractéristique. L'objectif du détecteur τ est de détecter si le vecteur fourni appartient ou pas à la macro-classe Ω_τ. Ce principe a été utilisé pour la détection de phones [1].
Les fonctions de sortie des détecteurs sont des fonctions *softmax*. Ceci permet d'obtenir une probabilité à posteriori d'appartenance à la macro-classe du vecteur caractéristique : $p(\Omega_\tau|\mathbf{a}_\tau)$. La macro-classification consiste à choisir l'expert qui fourni le maximum *a posteriori* (critère MAP) des probabilités (cf. Fig. 3) :

$$\Omega = \arg\max_{\Omega}\{p(\Omega_1|\mathbf{a_1}), \ldots, p(\Omega_\tau|\mathbf{a}_\tau)\} \qquad (13)$$

La macro-classification est facilitée par l'interaction entre les modèles DFE-NPC experts : chacun des experts a une "perception des autres".
Une fois la macro-classification achevée, le vecteur caractéristique est dirigé vers un classifieur.

5 Propriétés discriminantes du MER Coopératif

Dans cette section, nous démontrons les propriétés discriminantes du MER Coopératif par une modélisation statistique du processus. Le MER Coopératif est un mécanisme qui permet d'introduire une discrimination entre les classes d'une même macro-classe Ω_τ mais également avec les autres macro-classes $\Omega_{\overline{\tau}}$ (cf. équation 12) :

$$\begin{aligned}\Psi_i &= \frac{\sum_{j=1,j\neq i}^{M} L_j^i + L_{\Omega_{\overline{\tau}}}^i}{L_i^i} \\ &= \frac{\sum_{j=1,j\neq i}^{M} L_j^i}{L_i^i} + \frac{L_{\Omega_{\overline{\tau}}}^i}{L_i^i} \\ &= \Gamma_i^M + \Upsilon_i^{\Omega_{\overline{\tau}}}\end{aligned} \qquad (14)$$

Cette équation révèle deux termes :

– Un premier terme Γ_i^M qui est associé au MER entre la classe i et les autres classes appartenant à la même macro-classe. Ce terme de discrimination entre les classes est le MER "classique" qui est lié au critère MMI [5].
– Le second terme $\Upsilon_i^{\Omega_{\overline{\tau}}}$ est un terme de discrimination entre les macro-classes. C'est le rapport de l'erreur de prédiction du phonème i par le modèle correct (c'est-à-dire celui associé à la classe i) et celle issue du modèle représentant les autres macro-classes.

Afin de mieux comprendre le MER Coopératif, il est nécessaire de le décrire par son processus statistique équivalent.

5.1 Processus statistique

Supposons que le modèle du conduit vocal puisse-être approximé par la relation suivante :

$$y_{i,k} = F_{\mathbf{w},\mathbf{a}_i}(\mathbf{y}_{i,k}) + \epsilon_{i,k} \tag{15}$$

Où $\mathbf{y}_{i,k}$ est le contexte de prédiction du phonème i : $[y_{i,k-1}, y_{i,k-2}, ..., y_{i,k-\lambda}]^T$. $\epsilon_{i,k}$ un bruit distribué de manière indépendante et identique. Les ϵ sont distribués selon une loi normale $\mathcal{N}(0,\Sigma)$ avec une moyenne nulle et une matrice de covariance diagonale $\Sigma = \sigma \mathrm{I}$. La vraisemblance conditionnelle des observations $y_{i,k}$ est donnée par :

$$\begin{aligned} p(y_{i,k}|\mathbf{y}_{i,k},\mathbf{w},\mathbf{a}_i) &= p(\epsilon_{i,k}) = \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon_{i,k}^2}{2\sigma^2}} \\ &= \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(F_{\mathbf{w},\mathbf{a}_i}(\mathbf{y}_{i,k}) - y_{i,k}\right)^2}{2\sigma^2}} \end{aligned} \tag{16}$$

Et la vraisemblance jointe de la fenêtre d'analyse du signal $\mathcal{Y}_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,K}]$ issue du phonème i composé de K échantillons (dans l'hypothèse d'une distribution équiprobable) est :

$$p\left(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_i\right) = \frac{1}{(\sqrt{2\pi\sigma^2})^K} e^{-\frac{\sum_{k=1}^{K}\left(F_{\mathbf{w},\mathbf{a}_i}(\mathbf{y}_{i,k}) - y_{i,k}\right)^2}{2\sigma^2}} \tag{17}$$

$\mathbf{Y}_i = \{\mathbf{y}_{i,1} \ldots \mathbf{y}_{i,K}\}$ est la séquence de fenêtres de prédiction du phonème i.

L'équation (17) montre que la minimisation de l'erreur de prédiction L_i^i du phonème i par le modèle NPC-2 $F_{\mathbf{w},\mathbf{a}_i}$ équivaut à la maximisation de la vraisemblance jointe (cf. équation 17) :

$$\begin{aligned} \min_{\mathbf{w},\mathbf{a}_i} L_i^i &= \min_{\mathbf{w},\mathbf{a}_i} \sum_k^K \left(y_{i,k} - F_{\mathbf{w},\mathbf{a}_i}(\mathbf{y}_{i,k})\right)^2 \\ &\iff \max_{\mathbf{w},\mathbf{a}_i} p(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_i) \end{aligned} \tag{18}$$

Sous les précédentes hypothèses, la minimisation de l'erreur de prédiction est une approximation de la maximisation de la vraisemblance.

5.2 Influence du terme de discrimination entre les classes d'une même macro-classe

Le premier terme du MER Coopératif (14) est le i-MER (8) pour les M classes : Γ_i^M. Nous avons montré que la minimisation de l'inverse du MER consiste à maximiser l'information mutuelle (MMI)[5].

La minimisation de l'inverse du MER consiste à minimiser le rapport entre l'erreur de prédiction du modèle correct (L_i^i) et les erreurs de prédiction des autres modèles ($\sum_{j=1,j\neq i}^M L_j^i$).

Selon l'équation (18), la minimisation de l'inverse du MER (8) peut-être vue comme le processus statistique suivant :

$$\begin{aligned} \min_{\mathbf{w},\mathbf{a}_i,\mathbf{a}_j} \frac{1}{\Gamma_i} &= \min_{\mathbf{w},\mathbf{a}_i,\mathbf{a}_j} \frac{L_i^i}{\sum_{j=1,j\neq i}^M L_j^i} \\ \iff \max_{\mathbf{w},\mathbf{a}_i,\mathbf{a}_j} &\frac{p(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_i)}{\sum_{j=1,j\neq i}^M p(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_j)} \end{aligned} \tag{19}$$

Ce processus est équivalent à la Maximisation de l'Information Mutuelle (MMI) (de plus amples détails sont disponibles dans [5]) :

$$\min_{\mathbf{w},\mathbf{a}_i,\mathbf{a}_j} \frac{1}{\Gamma_i} \iff \max_{\rho_i} \mathrm{I} = \max_{\rho_i} \frac{p(\mathcal{Y}_i|F_i)}{\sum_{j=1}^M p(\mathcal{Y}_i|F_j)p(F_j)} \tag{20}$$

Le terme Γ_i^M introduit un critère discriminant entre la classe i et les autres classes de la même macro-classe.

5.3 Influence du terme de discrimination entre les macro-classes

Le terme de discrimination entre les macro-classes $\Upsilon_i^{\Omega_{\overline{\tau}}}$ est le rapport entre les erreurs de prédiction issues du modèle de la classe i et le modèle représentant les "autres macro-classes" :

$$\Upsilon_i^{\Omega_{\overline{\tau}}} = \frac{L_{\Omega_{\overline{\tau}}}^i}{L_i^i} \tag{21}$$

Sous les conditions du modèle statistique décrit précédemment, ce rapport est équivalent à :

$$\max \Upsilon_i^{\Omega_{\overline{\tau}}} = \frac{L_{\Omega_{\overline{\tau}}}^i}{L_i^i} \iff \max \frac{p(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_i)}{p(\mathcal{Y}_i|\mathbf{Y}_i,\mathbf{w},\mathbf{a}_{\Omega_{\overline{\tau}}})} \tag{22}$$

Ce résultat montre qu'il y a une discrimination explicite entre la classe i appartenant à la macro-classe Ω_τ et les macro-classes $\Omega_{\overline{\tau}}$. En effet, ce rapport de vraisemblance peut-être vu comme une distance NPC introduite dans [10] entre ces deux modèles. La distance NPC est une extension de la distance d'Itakura au domaine non linéaire [12].

Dans cette section, nous avons montré comment les modèles Coopératifs DFE-NPC sont optimisés pour la discrimination explicite entre les classes d'une même macro-classe, mais aussi avec les "autres macro-classes". Par conséquent, les modèles experts ont une "perception" des autres experts.

6 Conditions expérimentales

Cette section décrit l'application de l'architecture Cooperative Modular NPC (CMNPC) à la classification de phonèmes.

Les différents phonèmes sont extraits de la base Darpa-Timit, à partir de l'ensemble des locuteurs de la première région (New England) dans le but de produire un environnement multi-locuteurs. Chaque phonème est divisé, selon sa durée, en un nombre de fenêtres de dimension fixée à 256 échantillons avec un entrelacement de 128 échantillons. Le nombre d'exemples est fixé à 300 par classes.

Nous proposons de comparer les modèles NPC avec les méthodes de codage traditionnelles : LPC, MFCC et PLP. La dimension du vecteur code est fixée à 12.

Le classifieur utilisé pour estimer les performances de tous les codeurs est un perceptron mutlti-couches (MLP) avec 12 entrées (dimension du vecteur code), 10 neurones en couche cachée et autant de sorties que de classes de phonèmes. La loi d'apprentissage est une descente de gradient utilisant l'algorithme de rétropropagation. Les détecteurs sont des MLP avec 12 entrées, 10 neurones en couche cachée et deux sorties avec des fonctions soft-max. La loi d'apprentissage est une descente de gradient utilisant l'algorithme Levenberg-Marquardt.

7 Classification de phonèmes

Dans cette section, nous présentons les résultats en classification de phonèmes. Les taux de classification présentés sont ceux de la base de test (300 fenêtres par classe) et le classifieur est entraîné de la même manière pour les différents codeurs.

Le tableau 4 présente les taux de détection pour chacune des macro-classes. La détection consiste à classer les codes fournis par l'expert DFE-NPC en deux classes : appartenir ou pas à la macro-classe Ω_7. Les taux montrent qu'il est possible de détecter les macro-classes en utilisant cette méthode. Le taux de détection moyen est de 83.42%. Les meilleures taux sont obtenus pour les macro-classes composées de deux classes comme les voyelles centrales (93.33%), les semi-voyelles (89.66%) et les liquides (89.82%). Par contre, il y a une plus grande confusion entre les fricatives (78.17% et 77.46%) et les occlusives (82.85% et 78.21%). Une première justification est le fait que la modélisation de phonèmes non voisés est une tâche compliquée pour les modèles prédictifs comme le modèle NPC. En effet, les sons voisés résultent d'une vibration périodique des cordes vocales, et par conséquent le signal contient des composantes prédictibles. Par contre les sons non voisés sont produits sans vibration des cordes vocales, ils sont le plus souvent modélisés par des bruits. Par conséquent, la modélisation de ces sons par un modèle prédictif est une tâche beaucoup plus complexe. De plus, nous n'avons pas utilisé de discrimination intrinsèque entre les macro-classes comme par exemple le voisement. La discrimination serait plus efficace si elle était plus pertinente. En effet, le codeur DFE-NPC expert de la macro-classe des fricatives non voisés doit discriminer les autres macro-classes (parmi ces macro-classes les sons voisés et non voisés), alors que la discrimination est nécessaire seulement pour les sons non voisés.

Ω_1	Voyelles antérieures : ih ey eh ae	82.85%
Ω_2	Voyelles centrales : ah er	96.33%
Ω_3	Voyelles postérieures : uw uh ow aa	78.33%
Ω_4	Diphtongues : ay aw oy	83.54%
Ω_5	Semi-Voyelles : y w	89.66%
Ω_6	Liquides : l r	89.82%
Ω_7	Nasales : m n ng	80.42%
Ω_8	Occlusives (voisées) : b d g	82.85%
Ω_9	Occlusives (non voisées) : p t k	78.21%
Ω_{10}	Fricatives (voisées) : v z jh	78.17%
Ω_{11}	Fricatives (non voisées) : f s ch	77.46%
Ω	Moyenne	**83.42%**

TAB. 4 – Taux de détection

Les taux de classification pour chaque macro-classe sont présentés tableau 5, ils prennent en compte les scores de l'étape de macro-classification. Les scores les plus élevés sont obtenus pour les classes présentant le moins de perte en macro-classification.

Ω_1	Voyelles antérieures : ih ey eh ae	55.69%
Ω_2	Voyelles centrales : ah er	90.7%
Ω_3	Voyelles postérieures : uw uh ow aa	45.23%
Ω_4	Diphtongues : ay aw oy	58.37%
Ω_5	Semi-Voyelles : y w	84.28%
Ω_6	Liquides : l r	75.14%
Ω_7	Nasales : m n ng	54.05%
Ω_8	Occlusives (voisées) : b d g	65.45%
Ω_9	Occlusives (non voisées) : p t k	58.21%
Ω_{10}	Fricatives (voisées) : v z jh	61.57%
Ω_{11}	Fricatives (non voisées) : f s ch	59.89%
Ω	Moyenne	**64.41%**

TAB. 5 – Taux de classification de l'architecture CMNPC.

Les performances de l'architecture coopérative sont comparées à celles des codeurs classiques (cf. tab. 6). Le taux de classification est de 64.41%, il est meilleur de près de 12% par rapport aux méthodes de codage de l'état de l'art comme le codeur PLP (52.3%).

LPC	MFCC	PLP	CMNPC
48.3%	51.25%	52.3%	64.41%

TAB. 6 – Taux de classification pour l'ensemble des phonèmes.

Le taux de classification de notre précédente architecture [6] sur la même tâche est de 61.65%, il y a donc une amélioration relative de 5%. La principale amélioration réside dans le concept de la nouvelle architecture. La précédente

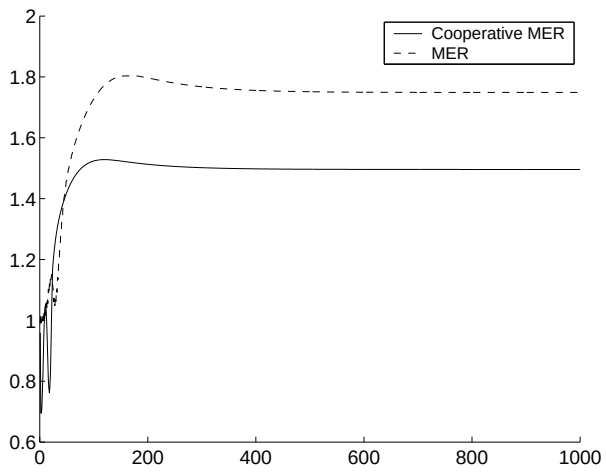

FIG. 4 – Evolution du MER et du MER Coopératif durant la phase de paramétrisation (voyelles antérieures)

est organisée autour d'un arbre de décision dont l'inconvénient majeur réside dans la profondeur de l'arbre. En effet, la profondeur de macro-classification des voyelles est plus grande que celle des phonèmes non voisés, les performances en sont donc réduites. Le concept parallèle de l'architecture coopérative permet de ne pas handicaper un type de phonème par rapport à un autre. De plus, l'interaction entre les modules permet à chaque expert d'avoir une bonne "perception" de l'espace d'entrée.

7.1 Influence de l'ajout d'une classe sur les taux de classification

Le processus de communication entre les codeurs que nous proposons consiste à ajouter systématiquement une nouvelle classe représentant les "autres macro-classes".
Le MER est une mesure de discrimination entre les modèles NPC. La figure 4 présente une comparaison de l'évolution du MER "classique" et du MER Coopératif pour le cas des voyelles antérieures. Le MER "classique" est beaucoup plus élevé que le MER Coopératif. Le fait d'ajouter une nouvelle classe a donc pour effet de diminuer la discrimination entre les modèles.
Pour mesurer l'influence de l'ajout de la classe, on paramétrise deux codeurs DFE-NPC avec et sans coopération pour les voyelles antérieures (formée de 4 classes de phonèmes : /ih/, /ey/, /eh/ et /ae/). Le modèle DFE-NPC "classique" est paramétrisé avec la base B1App (cf. tableau 7), alors que le modèle DFE-NPC Coopératif est paramétrisé avec la base B2App.
On teste les deux modèles uniquement sur la base B1Tst, ce qui a pour conséquence de mesurer les performances des experts uniquement sur leurs propres macro-classes. Chaque modèle code les exemples de la base B1Tst, ces codes sont ensuite dirigés vers un classifieur MLP de structure de 12 entrées (dimension du vecteur code), 10 cellules en couche cachée et 4 sorties pour les 4 classes. Le tableau

Nom	Nbre de classes
B1App	4 = les 4 voyelles
B1Tst	4 = les 4 voyelles
B2App	5 = les 4 voyelles + les autres macro-classes
B2Tst	5 = les 4 voyelles + les autres macro-classes
B3App	2 = les 4 voyelles & les autres macro-classes
B3Tst	2 = les 4 voyelles & les autres macro-classes

TAB. 7 – Constitution des bases pour la mesure de l'influence de la classe ajoutée

8 montre les taux de classification des voyelles antérieures. Du fait de la complexité accrue de la tâche consécutif à l'ajout d'une classe, les performances sont moins bonnes pour le modèle coopératif. Elles chutent de près de 4% entre le modèle coopératif et le modèle "classique" (cf. tableau 8).

DFE-NPC	Cooperative DFE-NPC
71.6%	67.22%

TAB. 8 – Taux de classification avec et sans coopération pour les voyelles antérieures

Cependant, la classe ajoutée est nécessaire pour la coopération entre les modules. En fait, elle permet d'introduire une meilleure "perception" de l'espace d'entrée par chacun des modules : chaque module a connaissance des exemples des "autres". Nous avons montré théoriquement comment le MER Coopératif (§5.3), introduit la discrimination explicite entre les macro-classes. Cette discrimination se retrouve dans le cas de la détection de macro-classes qui est en réalité une mesure de discrimination entre les macro-classes.
Pour cela, on mesure le taux de détection de la macro-classe des voyelles antérieures. Le test consiste à coder les exemples de la base B3App (cf. tableau 8), ensuite on entraîne un détecteur comme dans §4.3. Les mesures de généralisation se font sur la base B3Tst. Les différents modèles, DFE-NPC et DFE-NPC Coopératif codent les exemples. Le détecteur estime la probabilité d'appartenance ou pas à la macro-classe des voyelles antérieures. La prise de décision se fait en prenant le maximum de cette probabilité.
Le tableau 9 présente les taux de détection dans le cas des codeurs avec et sans classes ajoutées. On peut donc se rendre compte de l'intérêt de la classe ajoutée car il y a une différence notable entre les taux de détection ($\approx 20\%$). La classe ajoutée qui modélise les autres macro-classes a un intérêt certain dans le cas de la discrimination entre macro-classes. Le modèle coopératif atteint un score de détection de 82.85% alors que le modèle "classique" n'est que de 63.22%.
La méthode consistant à ajouter une classe qui modélise les autres macro-classes est donc une méthode de coopération simple mais efficace pour la discrimination entre les macro-classes. Cependant les performances pour chacune

DFE-NPC	Cooperative DFE-NPC
63.22%	82.85%

TAB. 9 – Taux de détection de la macro-classe voyelles antérieures (avec et sans coopération)

des macro-classes sont diminuées du fait de la complexité de la tâche de discrimination. Les performances globales montrent une nette augmentation du taux de classification sur l'ensemble des phonèmes.

8 Conclusions

Nous avons présenté une architecture coopérative pour l'extraction de caractéristiques discriminantes : Cooperative Modular NPC (CMNPC). Cette architecture est basée sur l'interaction entre des "experts" discriminants", les modèles DFE-NPC. Ces modèles fournissent la discrimination nécessaire par le biais d'un critère discriminant : le Rapport d'Erreur de Modélisation (MER). Ce critère est lié à la Maximisation de l'Information Mutuelle (MMI). Nous avons proposé une extension de ce critère pour la discrimination entre des groupes phonétiques (macro-classes). Ce critère coopératif est le Cooperative MER. Il permet à chaque expert d'avoir une meilleure "perception" des autres codeurs et donc de l'espace d'entrée. Par conséquent, les vecteurs codes fournis par chaque expert contiennent des caractéristiques discriminantes par rapport aux autres macro-classes. La macro-classification est basée sur des détecteurs qui fournissent des probabilités à priori d'appartenance aux différentes macro-classes. Une expertise phonétique permet de définir les différentes macro-classes afin d'optimiser le processus d'extraction de caractéristiques. De plus, nous avons montré de manière théorique et expérimentale que l'interaction entre les modèles DFE-NPC est une bonne méthode pour l'amélioration de la phase d'extraction de caractéristiques mais aussi pour la conception d'architectures modulaires discriminantes. Les tests montrent une amélioration de prés de 12% du taux de classification de phonèmes par rapport aux méthodes de codage traditionnelles. Cependant, les performances des experts sont affectées par l'ajout de la classe modélisant les autres macro-classes. Nos futurs axes de recherche concernent la coopération intelligente par l'utilisation, entre autres de connaissances phonétiques.

Références

[1] C. Antoniou. Modular neural networks exploit large acoustic context through broad-class posteriors for continous speech recognition. *Proc. of ICASSP*, 2001.

[2] G. Auda and M. Kamel. Cmm : Cooperative modular neural networks for pattern recognition. *Pattern Recognition Letters*, 18 :1391–1938, 1997.

[3] M. Birgmeier. Nonlinear prediction of speech signals using radial basis function network. *EUSIPCO*, 1996.

[4] H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18 :205–231, 1996.

[5] M. Chetouani, B. Gas, and J.L. Zarader. Maximization of the modelisation error ratio for neural predictive coding. *Proc. of NOLISP*, 2003.

[6] M. Chetouani, B. Gas, and J.L. Zarader. Modular neural predictive coding for discriminative feature extraction. *Proc. of ICASSP*, 2003.

[7] M. Chetouani, B. Gas, J.L. Zarader, and C. Chavy. Neural predictive coding for speech discriminant feature extraction : The dfe-npc. *Proc. of ESANN*, pages 275–280, 2002.

[8] O. Farooq and S. Datta. Phoneme recognition using wavelet based features. *Information Sciences*, 150 :5–15, March 2003.

[9] M. Faundez. *Modelado predictivo no lineal de la señal de voz aplicado a codificación y reconocimiento de locutor*. PhD thesis, Universitat politècnica de catalunya, 1998.

[10] B. Gas, J.L. Zarader, C. Chavy, and M. Chetouani. Discriminant features extraction by predictive neural networks. *Proc. of SSIP*, pages 64–68, September 2001.

[11] H. Hermansky. Should recognizers have ears ? *Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 1–10, 1997.

[12] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. ASSP*, 23(1) :67–72, February 1975.

[13] S. Katagiri. *Handbook of Neural Networks for Speech Processing*. Artech House eds., 2000.

[14] J.H. Lee, H.Y. Jung, T.W. Lee, and S.Y. Lee. Speech feature extraction using independent component analysis. *ICASSP*, 3 :1631–1634, 2000.

[15] Y. Liu, X. Yao, Q. Zhao, and T. Higuchi. An experimental comparison of ensemble learning on desicion boundaries. *Proc. of IJCNN*, pages 221–226, 2002.

[16] J. Mariani. *Analyse, synthèse et codage de la parole-Traitement automatique du langage parlé 1*. Hermés Science Pulications, 2002.

[17] L. Rabiner and B.J. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[18] S. Sivadas and H. Hermansky. Hierarchical tandem feature extraction. *Proc. of ICASSP*, 2002.

[19] H. Teager and S. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. *Proc. NATO ASI on Speech production and Speech Modeling*, pages 241–261, 1989.

[20] J. Thyssen, H. Nielsen, and S.D. Hansen. Nonlinearities short-term prediction in speech coding. *Proc. ICASSP*, 1 :185–188, 1994.

[21] S. Wang, E. Paksoy, and A. Gersho. Performance of nonlinear prediction of speech. *ICSLP*, 1 :29–32, November 1990.

[22] L. Wu, M. Niranjan, and F. Fallside. Fully vector quantized network-based code-excited predictive speech coding. *IEEE Trans. Speech and Audio Processing*, 2(4), October 1994.