Une architecture modulaire pour l'extraction de caractéristiques en reconnaissance de phonèmes

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 – Dans ce papier, nous présentons une architecture appelée Modular Neural Predictive Coding (MNPC). Elle est utilisée pour l'extraction de caractéristiques discriminantes. Cette architecture est conçue à l'aide de connaissances phonétiques. On estime les performances de cette architecture sur une tâche de reconnaissance de phonèmes extraits de la base Darpa-Timit. Une comparaison avec les méthodes de codage (LPC, MFCC et PLP) montrent une nette amélioration du taux de reconnaissance.

Abstract – In this paper, we present an architecture called the Modular Neural Predictive Coding Architecture (MNPC). The Modular NPC is used for Discriminative Feature Extraction (DFE). It provides an architecture based on phonetics knowledge applied to phoneme recognition. The phonemes are extracted from the Darpa-Timit speech database. Comparisons with coding methods (LPC, MFCC, PLP) are presented: they put in obviousness an improvement of the recognition rates.

1 Introduction

Dans l'objectif d'améliorer le processus de reconnaissance de la parole, plusieurs voies peuvent être choisies. Une d'elles est l'amélioration de l'étape d'extraction de caractéristiques. En effet, de récents travaux montrent l'importance de cette étape [5, 6, 9]. Elle est traditionnellement réalisée par des méthodes temporelles comme le codage LPC (Linear Predictive Coding), les méthodes fréquentielles comme le codage MFCC (Mel Frequency Cepstral Coding) ou bien les méthodes intégrant des connaissances sur la perception humaine comme le codage PLP (Perceptual Predictive Coding). Le problème avec ces méthodes est le manque de discrimination. Il n'y a pas de de mécanisme explicite permettant la discrimination entre les modèles.

La principale méthode pour l'introduction de la discrimination est l'extraction de caractéristiques discriminantes (Discriminative Feature Extraction DFE) basée sur la minimisation de l'erreur de classification (Minimum Classification Error MCE) [6]. L'idée clé de la méthode DFE est que les étapes extraction de caractéristiques et de classification doivent être entraînées simultanément afin d'améliorer le système de reconnaissance.

Une autre approche pour l'implémentation de la méthode DFE consiste à entraîner séparément l'extracteur de caractéristiques et le classifieur [4, 5]. Cette méthode est plus appropriée pour les problèmes complexes. En effet, durant l'apprentissage simultané de ces deux étapes, l'évolution des paramètres de l'extracteur de caractéristiques est plus petite que celle des paramètres du classifieur [5]. L'extracteur de caractéristiques doit être optimisés avec un critère qui mesure le pouvoir discriminant des caractéristiques sélectionnées. Par exemple, le critère de maximisation de l'information mutuelle (Maximization of the Mutual Information MMI) a été utilisé pour la sélection de caractéristiques [9].

Dans ce papier, on présente une architecture modulaire: Modular Neural Predictive Coding (MNPC) pour l'extraction de caractéristiques discriminante (DFE). Dans un premier temps, on présente le codage neuro-prédictif ainsi que son principe d'extraction de caractéristiques. Ensuite, un critère discriminant est introduit: le rapport d'erreur de modélisation. Ce critère est à la base de la définition d'une variante discriminante du codeur neuro-prédictif. La section 3 décrit la nouvelle architecture modulaire permettant la combinaison de plusieurs codeurs pour l'extraction de caractéristiques. Les sections suivantes présentent les conditions expérimentales et les résultats en reconnaissance de phonèmes. Et, finalement nous donnerons quelques conclusions sur le travail proposé.

2 Codage neuro-prédictif

Le codage neuro-prédictif (Neural Predictive Coding) [3] 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). Cette stratégie est consistante avec le fait que le processus de production de la parole est connue pour être non linéaire [8].

Les signaux de parole sont divisés en un nombre fixe de fenêtres et l'échantillon courant est prédit par une combinaison des échantillons passés. λ est la dimension de la fenêtre de prédiction :

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

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

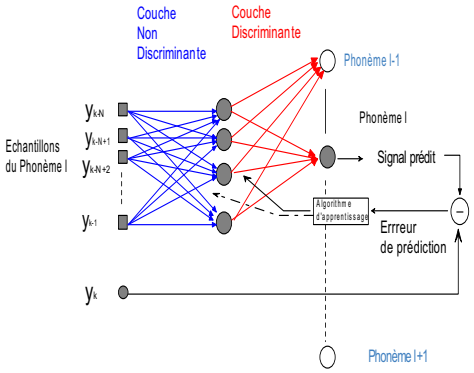

FIG. 1: Architecture du codeur NPC

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 $\mathbf{z}_k = G_{\mathbf{w}}(\mathbf{y}_k)$.

La fonction F est un modèle auto régressif non linéaire du signal de parole. Cette modélisation est une extension du modèle auto régressif linéaire utilisé par le codage linéaire prédictif (LPC).

Un des problèmes avec une approche par réseaux de neurones prédictifs est que cela génére un grand nombre de paramètres (les poids du réseau). Le but est de limiter ce nombre, et l'idée clé du modèle NPC est de permettre d'avoir un nombre arbitraire de coefficients indépendamment de la dimension de la fenêtre de prédiction. La solution consiste à considérer les poids de la seconde couche $\mathbf{a}$ comme les coefficients. Ces poids doivent modéliser le phonème. Cela est obtenu par la détermination des poids de la couche cachée $\mathbf{w}$ par tous les échantillons de tous les phonèmes tandis que la seconde couche est spécifique à chaque phonème:

La phase d'apprentissage du codeur NPC s'effectue en deux étapes:

- *La phase de paramétrisation*. Cette phase consiste à déterminer tous les poids de la première couche $\mathbf{w}$ qui sont les paramètres du codeur.
- *La phase de codage*. 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.

2.1 Le codage NPC-2: notion de classes durant la paramétrisation

Le modèle NPC-2 est une extension du modèle NPC qui permet d'incorporer la notion de classes durant la phase de paramètrisation. Ceci est fait en limitant les poids de la couche de sortie à un vecteur code par classe de phonèmes au lieu d'un vecteur par phonème.

Pour les M classes possibles et C_i classe du phonème i, la fonction coût du modèle NPC-2 est:

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

Le symbole de Kronecker associe la classe C_i à la couche de sortie l.

Une fois cette fonction coût minimisée, l'encodeur est prêt à coder les données. La phase de codage est identique au modèle NPC.

2.2 Extraction de caractéristiques

L'objectif du modèle NPC est de pouvoir réaliser des poids de couches de sorties discriminants. Ces poids $\mathbf{a}_l$, les vecteurs code, doivent représenter des caractéristiques phonétiques discriminantes. Les poids $\mathbf{a}_l$ de la première couche sont communs à tous les phonèmes. Cette stratégie est consistante avec le fait que le modèle de production de la parole peut-être séparer en une partie commune (le conduit vocal) et des zones spécifiques (les contributions glottales et les radiations) [7].

Considérons i et j deux phonèmes appartenant respectivement aux classes différentes C_i et C_j. Les modèles NPC-2 associés à ces deux phonèmes sont les suivants:

$$\begin{cases} F_{\mathbf{w},\mathbf{a}_i} = H_{\mathbf{a}_i} \circ G_{\mathbf{w}} \\ F_{\mathbf{w},\mathbf{a}_j} = H_{\mathbf{a}_j} \circ G_{\mathbf{w}} \end{cases} \tag{4}$$

Les modèles NPC-2 $F_{\mathbf{w},\mathbf{a}_i}$ et $F_{\mathbf{w},\mathbf{a}_j}$ sont différents alors $G_{\mathbf{w}}$ est commune aux deux phonèmes i et j. Cette fonction rassemble les caractéristiques communes alors que les caractéristiques discriminantes sont représentées par les fonctions $H_{\mathbf{a}_i}$ et $H_{\mathbf{a}_j}$.

2.3 Le rapport d'erreur de modélisation

Le principal problème avec les approches prédictives est le manque de discrimination: les modèles sont entraînés de manière indépendantes. Par conséquent, il n'y a pas de discrimination explicite entre les modèles. Afin de résoudre ce problème nous avons développé une mesure de discrimination entre les modèles NPC-2: le Rapport d'Erreur de Modélisation (MER) [3].

Soit L_j^i l'erreur de prédiction du phonème i issue du modèle NPC-2 $H_{\mathbf{a}_j}$ qui est associé au phonème j:

$$L_j^i = \sum_k (y_{i,k} - H_{\mathbf{a}_j} \circ G_{\mathbf{w}}(\mathbf{y}_{i,k}))^2 \tag{5}$$

Le i-MER est le rapport inverse de l'erreur de prédiction du phonème i par le modèle NPC-2 correct par les erreurs de prédiction du même phonème i par les autres modèles NPC-2:

$$\Gamma_i = \frac{\sum_{j=1, j\neq i}^{M} L_j^i}{L_i^i} \tag{6}$$

Où M est le nombre de classes.

La maximisation du i-MER est utilisée pour la discrimination entre des modèles NPC-2. En effet, nous avons montré comment il est lié au critère de maximisation de l'information mutuelle (MMI) [1].

En accord avec la définition du i-MER (6), on étend le MER à toutes les M classes:

$$\Gamma = \frac{Q^d}{(M-1)Q^m} \tag{7}$$

Avec $Q^d = \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} L_j^i$ et $Q^m = \sum_{i=1}^{M} L_i^i$.

La maximisation du rapport d'erreur de modélisation permet une extraction de caractéristiques discriminantes: le DFE-NPC [3]:

$$Q_{DFE-NPC} = \frac{1}{\Gamma} \tag{8}$$

La loi de modification des poids a est proportionnelle à l'inverse du gradient du MER $Q_{DFE-NPC}$ (8):

$$\frac{\partial}{\partial a}\left(\frac{1}{\Gamma}\right) = \frac{M-1}{Q^d}\left(\frac{\partial Q^m}{\partial a} - \frac{1}{\Gamma}\frac{\partial Q^d}{\partial a}\right) \tag{9}$$

Le modèle DFE-NPC possède les propriétés nécessaires pour l'extraction de caractéristiques discriminantes. Il a été appliqué avec succès à la reconnaissance d'un nombre limité de phonèmes [3]. Dans cet article, 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 nombre différent de phonèmes est de trouver ceux qui a de commun entre eux. La modularité est une des solutions pour résoudre ce problème car elle permet de combiner experts pour chaque type de phonème.

3 Modular NPC

L'extraction de caractéristiques peut être améliorer en suivant la méthode "diviser pour mieux régner": un problème complexe est divisé en sous-problèmes plus simples. Traditionnellement, l'extraction de caractéristiques est réalisée de la même manière pour tous les phonèmes en dépit de leurs différences. En effet, il existe plusieurs sortes de différences comme le voisement par exemple. L'idée clé de l'architecture Modular NPC est de permettre d'obtenir un traitement adapté à chaque catégorie de phonèmes. Les catégories ou macro-classes sont extraites de l'Alphabet Phonétique Internationale (API). Les phonèmes d'une même catégorie ont des caractéristiques proches.

3.1 Description

L'architecture Modular NPC (MNPC) [2] est composée de deux étapes: une étape de macro-classification et une autre de codage par des DFE-NPC experts. La macro-classification permet de rediriger le phonème vers l'expert adéquate. Le phonème sera ensuite codé par un DFE-NPC expert dans le traitement d'une macro-classe API. Les experts sont donc optimisés par le biais de la maximisation du MER sur une seule macro-classe.

3.2 L'étape de marco-classification

La macro-classification se fait par le biais d'un arbre de décision (cf. Tab. 1). Le phonème est guidé de nœud en nœud vers un expert. Les macro-classifieurs sont entraînés non pas avec une notion de classe comme les codeurs NPC-2 ou DFE-NPC mais avec une notion de macro-classe. Le critère de discrimination entre ces macro-classes est la maximisation du rapport d'erreur de modélisation (MER).

Pour un macro-classifieur τ dont la fonction est de discriminer entre les $\mathbf{\Omega}$ macro-classes, la fonction coût est définie comme suivant:

$$L = \sum_i \sum_k \sum_l (y_{i,k} - \Phi_{\mathbf{w},\mathbf{a}_{\Omega_l}}(\mathbf{y}_{i,k}))^2 \delta_{\Omega_i - l} \tag{10}$$

Ω_i est la macro-classe du phonème i. $\Phi_{\mathbf{w},\mathbf{a}_l}$ est une des $\mathbf{\Omega}$ fonctions.

Contrairement au modèle NPC, les codes résultant de la phase de *paramétrisation* sont utilisés pour la classification. La macro-classification est effectuée par une méthode prédictive:

$$\Omega_i = \arg\min_{\Omega} \sum_k \sum_l (y_{i,k} - \Phi_{\mathbf{w},\mathbf{a}_{\Omega_l}}(\mathbf{y}_{i,k}))^2 \tag{11}$$

Après l'étape de macro-classification, le phonème est orienté vers "DFE-NPC expert" qui permet de fournir le vecteur code représentant de ce phonème.

TAB. 1: Description de l'architecture Modular NPC

Macro-Classifieur	Nœud	Classes
Niveau 1	1	Voisés / non voisés
Niveau 2	1	Voyelles / Consonnes
	2	Oclusives / Fricatives (non voisés)
Niveau 3	1	Voyelles-Diphtongues / Semi-Voyelles
	2	Nasales-Liquides / Oclusives-Fricatives (voisés)
Niveau 4	1	Voyelles/Diphtongues
	2	Nasales /Liquides
	3	Oclusives/Fricatives (voisés)
Niveau 5	1	Antérieures/ Centrales/ Postérieures Voyelles

4 Conditions expérimentales

Cette section décrit l'application de l'architecture Modular NPC à la reconnaissance de phonèmes.

Les différents phonèmes sont extraits de la base Darpa-Timit. Les phonèmes sont extraits 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 (en apprentissage et en test).

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. Le classifieur est entraîné de la même manière pour les différents codeurs, et on présente les taux de reconnaissance de la base de test.

5 Reconnaissance de phonèmes

Dans cette section, nous présentons les résultats en reconnaissance de phonèmes.

Le principe d'extraction de caractéristiques de l'architecture MNPC est composée de deux phases: la macro-classification et le codage. La macro-classification joue un grand rôle. Les DFE-NPC sont optimisés dans l'objectif d'être expert dans la discrimination entre les classes d'une même macro-classe. Par conséquent, si le phonème est présenté au codeur qui n'est pas adapté à son traitement alors les performances vont décroître.

Les performances des macro-classifieurs sont présentées Tab. 2. Les taux de reconnaissance montrent une difficulté de macro-classification des phonèmes voisés et entre autres les voyelles. Les profondeurs de l'arbre sont différentes en fonction de la macro-classe traitée. Les meilleures performances sont obtenues pour les macro-classes de phonèmes non voisés.

TAB. 2: Taux de reconnaissance des macro-classifieurs

Voisés / non voisés	98.74%
Voyelles / Consonnes	83.3%
Oclusives/Fricatives (non voisés)	98.33%
Voyelles- Diphtongues /Semi-Voyelless	82.3%
Nasales-Liquides/ Oclusives-Fricatives (voisés)	93.03%
Voyelles/Diphtongues	88.4%
Nasales/ Liquides	96.14%
Oclusives / Fricatives (voisés)	95.28%
Antérieures/ Centrales/ Postérieures	77.3%

La disparité de profondeur de l'arbre en fonction des macro-classes affecte directement les performances finales. Le tableau 3 présente les résultats finaux pour les différentes classes en prenant en compte les scores de l'étape précédente: la macro-classification. Les meilleurs résultats sont obtenus pour les phonèmes non voisés (occlusives 88.99% et fricatives 77.43%). Ces résultats sont comparables à ceux obtenus pour les phonèmes voisés: occlusives 72.94% et fricatives 70.65%. Les résultats pour les voyelles sont moins bon du fait des taux de macro-classification.

TAB. 3: Taux de reconnaissance pour chaque classe

Ω_1	Voyelles antérieures: ih ey eh ae	39.72%
Ω_2	Voyelles centrales: ah er	41.45%
Ω_3	Voyelles postérieures: uw uh ow aa	36.08%
Ω_4	Diphtongues: ay aw oy	56.64%
Ω_5	Semi-Voyelles: y w	64.65%
Ω_6	Liquides: l r	75.46%
Ω_7	Nasals: m n ng	57.61%
Ω_8	Occlusives (voisés): b d g	72.94%
Ω_9	Occlusives (non voisées): p t k	88.99%
Ω_{10}	Fricatives (voisés): v z jh	70.65%
Ω_{11}	Fricatives (non voisées): f s ch	74.43%

Le taux de reconnaissance globale est de 61.65% (cf. Tab. 4). Cependant, dans les mêmes conditions, on montre une nette amélioration (environ 10%) du taux de reconnaissance par rapport aux méthodes traditionnelles LPC, MFCC et PLP (cf. Tab. 4).

TAB. 4: Taux de reconnaissance pour la globalité des phonèmes

LPC	MFCC	PLP	Modular NPC
48.3%	51.25%	52.3%	61.65%

6 Conclusions

Nous présentons une architecture pour l'extraction de caractéristiques discriminantes (Discriminative Feature Extraction DFE): the Modular Neural Predictive Coding (MNPC). Elle est basée sur la combinaison de deux phases: macro-classification et codage. La macro-classification est réalisée par un arbre de décision où chaque nœud un classifieur prédictif. Cette arbre de décision est conçue à l'aide de connaissances phonétiques. Les DFE-NPC experts permettent d'obtenir la discrimination nécessaire à la tâche par le biais de la maximisation du rapport d'erreur de modélisation (MER). Les résultats expérimentaux montrent une amélioration nette du score de reconnaissance de phonèmes, environ 10% par rapport aux méthodes de codage traditionnelles.

Références

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

[2] M. Chetouani, B. Gas, and J.L. Zarader. Modular neural predictive coding for discriminative feature extraction. *Proc. of ICASSP*, 2:33–36, 2003.

[3] 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.

[4] A. de la Torre, Antonio Peinado, Antonio J. Rubio, Victoria E. Sánchez, and Jesús E. Diaz. An application of minimum classification error to feature space transformations for speech recognition. *Speech Communication*, 20:273–290, 1996.

[5] A. de la Torre, Antonio Peinado, Antonio J. Rubio, José C. Segura, and C. Benitéz. Discriminative feature weighting for hmm-based continous speech recognizers. *Speech Communication*, 38:267–286, 2002.

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

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

[8] 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.

[9] K. Torkkola. On feature extraction by mutual information maximization. *ICASSP*, 1:821–824, 2002.