ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 275-280

Neural Predictive Coding for Speech Discriminant Feature Extraction : The DFE-NPC

M. CHETOUANI, B. GAS, J.L. ZARADER, C. CHAVY
Laboratoire des Instruments et Systèmes d'Ile de France (LISIF)
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 chavy@ccr.jussieu.fr

Abstract : In this paper, we present a predictive neural network called Neural Predictive Coding (NPC). This model is used for non linear discriminant features extraction (DFE) applied to phoneme recognition. We also, present a new extension of the NPC model : DFE-NPC. In order to evaluate the performances of the DFE-NPC model, we carried out a study of Darpa-Timit phonemes (in particular /b/, /d/, /g/ and /p/, /t/, /q/ phonemes) recognition. Comparisons with coding methods (LPC, MFCC, PLP, RASTA-PLP) are presented: they put in obsviousness an improvement of the classification.

1. Introduction

Significant advances have been made in the area of speech coding over the last decade. Most of them have been done in the context of speech recognition. In fact, coding methods allow to improve recognition rate by extracting discriminant features from speech signals. Linear predictive coding (LPC) is known as a useful method for feature extraction based on a modelisation of the vocal tract. In frequential domain, the most often used is the mel frequency cepstral coding (MFCC) using the Mel scale which is based on the human ear scale. Most of the recent works aiming at improving the discriminant features extraction are frequency-based [1,2].

In the temporal domain, predictive methods have been essentially developed in two ways. First, by introducing an estimation of the auditory properties of human ear, like in Perceptual Linear Prediction (PLP) [3]. An improvement of this method, consists in a special filtering : The Relative Spectral Technique (RASTA-PLP) [4].

Another strategy, is to develop nonlinear predictor [5,6]. In fact, it is known that the modelisation of the vocal tract is nonlinear [5] so nonlinear extraction can get an improvement of the recognition rate.

The implementation of nonlinear predictors is essentially based on two techniques ; Volterra filters [7] and neural networks [8,9]. The major advantage of Volterra filters is that, like in linear predictors, the least mean square solution for the filter coefficients can be expressed analytically. The main drawback lie in the fact that the number of coefficients grows fast with the prediction window. Predictive neural networks have already been successfully applied to speech processing [6]. But same drawbacks (the number of coefficients) occurs with neural networks. In addition, the weigths solution cannot be expressed analytically.

The NPC model [10] has the major advantage to allow an arbitrary limited number of coding coefficients. It has already been extended by incorporating class informations to improve the next pattern recognition stage, the NPC2- model [11,12]. In our paper, we focuse on en extension of the Neural Predictive Coding called DFE-NPC model applied to nonlinear dicriminant features extraction.

The paper is organized as follows : In section 2, we describe the NPC-2 model. We then present an extension of the NPC-2 model for discriminative features extraction (DFE) : the DFE-NPC model. We also compare the DFE-NPC model with traditionnal coding methods on phoneme recognition task.

2. The NPC-2 model

The Neural Predictive Coding [11, 12] model is an extension of the LPC traditional coding (Linear Predictive Coding) to the modelling of non linear signals. It is based on a two-layer perceptron composed of one hidden layer followed by one output cell: the *prediction cell* (see figure 1).

For such a task, the speech signal is divided into fixed length frames and the current speech sample is predicted from a combination of finite past samples. L being the length of the *prediction window*, one has:

$$\hat{y}_k = \mathrm{F}(\mathbf{y}_k) \text{ with } \mathbf{y}_k = \left[y_{k-1}, y_{k-2}, \cdots y_{k-L}\right]^{\mathrm{T}} \tag{1}$$

F is a non linear function which is composed of two functions $\mathrm{G}_{\mathbf{w}}$ (corresponding to the hidden layer) and $\mathrm{H}_{\mathbf{a}}$ (corresponding to the output layer) :

$$\mathrm{F}_{\mathbf{w},\mathbf{a}} = \mathrm{H}_{\mathbf{a}} \circ \mathrm{G}_{\mathbf{w}} \quad \text{with } \hat{y}_k = \mathrm{H}_{\mathbf{a}}(\mathrm{z}_k) \text{ and } \mathrm{z}_k = \mathrm{G}_{\mathbf{w}}(\mathbf{y}_k) \tag{2}$$

$\mathbf{w}$ denotes the hidden layer weights vector and $\mathbf{a}$ the output layer weights vector. All network weights are usually computed by minimizing a prediction cost function as the quadratic error criterion : $\mathrm{L} = \sum_k (y_k - \hat{y}_k)^2$ (3)

Where k is the index of the phoneme samples

For instance, over all the samples composing a signal, one can obtain after learning a function F which is a nonlinear auto regressive model (NLAR) of the signal.

One problem that occurs with this approach is that it generates a great number of parameters. The aim is to limit this number, and the key idea of the NPC coding is to allow an arbitrary number of coding coefficients by creating a second layer for each class phoneme, the first layer remaining the same for all classes. The cost function previously defined becomes : $\mathrm{L} = \sum_i \sum_k \sum_l \left(y_{i,k} - \mathrm{F}_{\mathbf{w},\mathbf{a}_l}(\mathbf{y}_{i,k})\right)\delta_{c_i - l}$ (4)

C_i is the class membership of phoneme *i* among a set of *M* possible classes. $\mathrm{H}_{\mathbf{a}_l}$ is one of the M functions corresponding to the $\mathbf{a}_l$ output layer weights. δ is the Kronecker symbol which associates the class C_i to the output layer *l*.

The learning process needs to be broken down in the following two phases : the *parameters adjustment phase* and the *coding phase.* During, the *parameters adjustment phase,* all the network weigths are estimated from a learning set composed of phonemes belonging to the M classes. Next, the output layers weigths are no longer used while the hidden layer weigths become the encoder *parameters*. In a second time, during the *coding phase*. The network works as a two layers perceptron composed of the hidden layer previously computed and of one output cell. These weigths are the only ones requiring updating. They are the NPC coding coefficients.

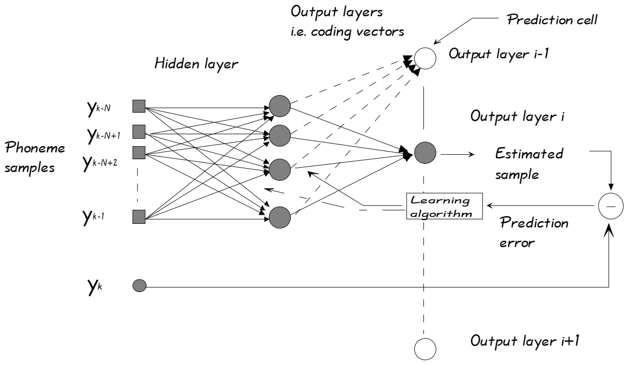

Figure 1 Architecture of the NPC

3. DFE-NPC model

To guarantee a class-discriminant features extraction, one can add constraints to the weigths evolution during the learning process. One possible mechanism is to introduce explicit discrimination between models. After the parameters adjustment phase of a phoneme i, we obtain the $\mathrm{F}_{\mathbf{w},\mathbf{a}_i}$ NPC model. To get discrimination, one can estimate the prediction error of the phoneme i using another model which is the $\mathrm{F}_{\mathbf{w},\mathbf{a}_j}$ model.

As a result, the prediction error is: $L_j^i = \sum_k \left(y_{i,k} - \mathrm{F}_{\mathbf{w},\mathbf{a}_j}\left(\mathbf{y}_{i,k}\right)\right)^2$.

The mechanism which discourage the models from resembling each other is obtained by the maximisation of the modelisation error ratio (MER) [12], Γ_{NPC} :

$$\Gamma_{NPC} = \frac{Q^d}{(M-1)Q^m} \qquad \text{with} \qquad Q^d = \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \mathrm{L}_j^i \qquad \text{and} \qquad Q^m = \sum_{i=1}^{M} \mathrm{L}_i^i .$$

(9)

Where Q^m is the NPC-2 model prediction cost function which has to be minimized, while the Q^d discriminant cost function which has to be maximized. So it is possible to optimise both the discrimination and the prediction error by applying the criterion which consists in the minimisation of the reverse MER : $Q_{NPC3} = \frac{1}{\Gamma_{NPC}}$ (10)

The modification law of any a or w weights is proportionnal to the gradient of Q_{NPC3} :

$$\frac{\partial}{\partial a}\left(\frac{1}{\Gamma_{NPC}}\right) = \frac{M-1}{Q^d}\left(\frac{\partial Q^m}{\partial a} - \frac{1}{\Gamma}\frac{\partial Q^d}{\partial a}\right) \tag{11}$$

It is composed of two terms, the first term corresponding to the prediction error minimisation (The NPC-2 cost function) and the second to the discrimination measure maximisation.

4. Experimental conditions

To evaluate the NPC performances we tested it in a phoneme recognition task. We will describe in this part the experimental conditions.

4.1. The database

We built several phonemes bases extracted from the Darpa-Timit [13] database. This database is composed of speakers speaking 10 different dialects of the United States. The evaluation is done on 3 bases. The first base groups four classes of voiced phonemes (vowels) and is constituted of 500 examples. The two other bases (constitued of 100 examples) : /b/, /d/, /g/ (voiced plosives) and /p/, /t/, /q/ (unvoiced plosives) are particularly interesting because they are frequently used and their identification is considered to be difficult. Those phonemes have been used by Lang and Waibel [14] to validate the Time Delay Neural Network (TDNN). The phonemes are chosen randomly among all available speakers to produce a multi-speaker environment. Every phoneme according to its duration, is divided into windows of a fixed length (256 samples), each of them being a phoneme example.

4.2. Traditional coding methods

Our aim is to test the efficiency of the speech features extraction of the DFE-NPC encoder. The performance will be estimated by classification. We made comparisons between NPC coding and traditional coding methods; LPC, MFCC coding and perceptual methods : PLP coding (Perceptual Linear Predictive coding) and RASTA-PLP coding (Relative Spectral Technique PLP coding). The number of coding coefficients is set to 12.

4.3. Classification with MLP

The classifier used to estimate performances of coding method is a basic MLP with 12 inputs (coding vectors dimension), 10 hidden neurons and as many outputs as there are phoneme classes. The learning rule is the gradient descent using error back propagation algorithm.

4.4. NPC evaluation using MLP classifier

In this paragraph, we present the results of classification of the different phoneme bases using the different coding methods; NPC-1, NPC-2 and DFE-NPC, and the traditional methods; LPC, MFCC, PLP and RASTA-PLP. We measure the generalisation score .

On table 1, one can see comparisons between recognition rates obtained by MLP classifier for vowels. Recognition rates have been obtained after 30000 learning iterations. The NPC coders give better results in generalisation, and one can see the better performance of DFE-NPC.

LPC	MFCC	PLP	RASTA-PLP	NPC-1	NPC-2	DFE-NPC
56.23%	58.25%	57%	59.25%	61%	62.95%	65.25%

Table 1 Recognition rates obtained with MLP classifier for /aa/, /ae/, /ey/, /ow/ phonemes

One can note on table 2, that the results for /b/, /d/, /g/ phonemes are consistent with the fact that they are voiced phonemes. Phonemes /p/, /t/, /q./ are unvoiced phonemes, so spectral method like MFCC or perceptual methods like PLP or RASTA-PLP have better performances than predictive methods like LPC, NPC-1 and NPC-2. But, one can note on table 2 that the discrimination introduced in the DFE-NPC encoder, compensates the default of temporal encoders for unvoiced phonemes. They are less predictable than voiced phonemes.

	LPC	MFCC	PLP	RASTA-PLP	NPC-1	NPC-2	DFE-NPC
/b/-/d/-/g/	57.28%	62%	64%	64%	65%	70.4%	73%
/p/-/t/-/q/	62.3%	66.6%	66%	68.33%	63.33%	65%	70.3%

Table 2 Recognition rates obtained with MLP classifier for /b/-/d/-/g/ and /p/-/t/-/q/ phonemes

The non linear features present in speech signals are better taken into account by the NPC-1, NPC-2 and DFE-NPC models. Moreover, the discriminant optimisation gives better results for unvoiced phonemes. The DFE-NPC encoder allows discrimant features extraction.

A comparison between the NPC-2 model and DFE-NPC model, shows the better discrimination of the DFE-NPC model. In fact, the DFE-NPC MER is higher than the NPC2 MER (see figure 2). Moreover, a study of the between-class covariance show that the DFE-NPC between class covariance converges to a higher value than NPC-2.

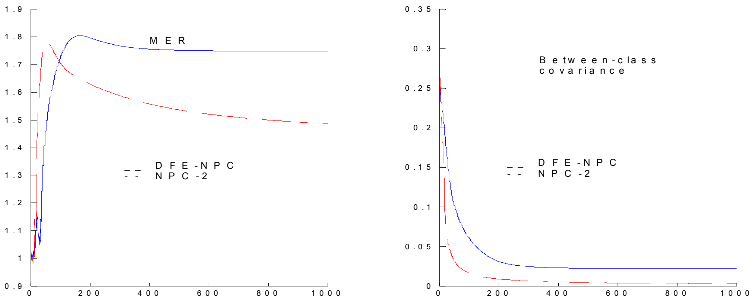

Figure 2 Modelisation Error Ratio evolution and between class covariance (vowels)

5. Conclusions

We have presented a non linear coding model to code speech signals ; the NPC model. This model, with the new optimisation (DFE-NPC model), allows discriminant features extraction (DFE). Both the discrimination and the prediction error are optimised, this optimisation discourages phoneme models from resembling each other. This model is compared with traditionnal and perceptual coding methods. We showed a significant improvement of the recognition rate specially in the case of unvoiced phonemes. Thanks to the MER, one can measure the discriminant properties of the encoder. Consequently, one can stop the *parameters adjustment phase* when the discrimination is optimal.

References:

[1] B.H. Juang and S. Katigiri "Discriminative Learning for Minimum Error Classification", IEEE Trans. On Signal Processing, Vol. 40, n°12, pp. 3043-3054, 1992.

[2] A. Biem, S. Katigiri, E. McDermott & B.-H. Juang *"*An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition*",* IEEE Trans. SAP, Vol. 9, pp. 96-110, 2001.

[3] H. Hermansky "Perceptual Linear Predictive (PLP) analysis of speech", Journal of the Acoustical Society of America, Vol. 4, pp. 1738-1752, 1990.

[4] H. Hermansky, N. Morgan "RASTA Processing of Speech", IEEE Trans. On Speech and Audio Processing, Vol. 2, pp. 587-589, October 1994.

[5] B. Townshend "Nonlinear Prediction of Speech", ICASSP'91, Vol.1, pp. 425-428, 1991.

[6] M. Faundez, E. Monte, F. Vallverdu "A Comparative Study Between Linear And Nonlinear Speech Prediction", Biological & artificial computation: from neuroscience to technology. IWANN'97, pp. 1154-1163, 1997.

[7] P. Chevalier, P. Duvaut, B. Pincinbone "Le Filtrage de Volterra Transverse Réel et Complexe en Traitement du Signal", Traitement du Signal, Vol. 7, n°5, pp. 451-476, 1990.

[8] A. Hussain "Locally-Recurrent Neural-Networks For Real-Time Adptatve Nonlinear Prediction of Non-Stationary Signals", Control And Intelligent Systems, Vol. 28, pp. 64-71, 2000.

[9] G. Dreyfus, O. Macchi, S. Marcos, O. Nerrand, L. Personnaz, Roussel-Ragot, D. Urbani and C. Vignat "Adaptive Training of Feedback Neural Networks for Non Linear Filtering", Neural Networks for Signal Processing, Vol. 2, pp. 550-559, 1992.

[10] B. Gas, J.L. Zarader, C. Chavy *"*A New Approach to Speech Coding : The Neural Predictive Coding"*,* Journal of Advanced Computational Intelligence, Vol.4, pp. 120-127, January 2001.

[11] C. Chavy, B. Gas, J.L. Zarader "Discriminative Coding with Predictive Neural Networks", ICANN'99, pp. 216-220, 1999.

[12] B. Gas, J.L. Zarader, C. Chavy, M. Chetouani *"*Discriminant Features Extraction by Predictive Neural Networks*",* WSES Inter. Conf. In Signal, Speech and Image Processing, Advances in Signal Processing and Communications, Malta, pp. 64-68, September 2001.

[13] University of Pennsylvania Linguistic Data Consortium. Darpa-timit: a multi speakers data base.

[14] K.J. Lang, A.H. Waibel and G.E. Hinton. *"*A Time-Delay Neural Network Architecture for Isolated Word Recognition*".* Neural Networks*,*Vol. 3, pp 23-43, 1990.