GLOBAL DISCRIMINATION FOR NEURAL PREDICTIVE SYSTEMS BASED ON N-BEST ALGORITHM.

Abdelhamid MELLOUK(*), Patrick GALLINARI(**)

(*) LRI, UA 410 CNRS, Bat 490
Université Paris-Sud
91405 Orsay Cedex
France
mellouk@lri.lri.fr

(**) LAFORIA, UA CNRS 1095
Université Paris 6
4 place Jussieu
75252 Paris Cedex 05 France
gallinari@laforia.ibp.fr

ABSTRACT

We describe a general formalism for training neural predictive systems. We then introduce discrimination at the frame level and show how it relates to maximum mutual information training. Last, we propose an approach for performing discrimination in predictive systems at the sequence level, it makes use of N-Best sequence selection. Performances for acoustic-phonetic decoding reach 77.4% phone accuracy on 1988 version of TIMIT.

1. INTRODUCTION

Recently, several hybrid models combining Neural Networks (NN) and dynamic programming segmentation have been proposed in the hope of improving continuous speech recognition systems. These systems differ by the organisation of the overall recognizer and the goal NN are used for: probability estimation, signal production or pre-processing of the data. Neural predictive systems (NPS) for continuous speech recognition (CSR) are one of these approaches [1,2,3]. Although early performances of these systems were rather disappointing, they have undergone recently several improvements and offer now good performances, while remaining easy to implement. All the models which have been proposed in the literature share the same basic scheme :

- predictive neural networks are used for low level modelling of words or phonemes. For the latter, one phonemic model, built from a small number of predictors -or states-, is trained for each phoneme to predict frame at time t from context frames.

- Speech signal is processed in parallel by all the models, giving one score per frame and state model. Sequences of matching scores between models computed outputs and reference templates are then processed through dynamic programming (DP) for segmentation. For training, the error corresponding to the best path is back propagated through the prediction models. This process is iterated until convergence. For recognition, only the forward pass of the above procedure is used.

These non-linear predictors make only local stationary assumptions about data which may be convenient for modelling speech [4], they allow to take easily into account context and to incorporate speech dynamic information.

We present in 2. a general formalism for training Neural Predictive Systems (NPS). In 3 we describe how to introduce discrimination at the frame level and relate this to Maximum Mutual Information training. In 4 we introduce phone level discrimination through N-Best segment selection.

2. TRAINING NPS

Our NPS [3] makes use of three predictive multi-layer perceptrons -or state models- for modelling each phoneme. Speech is considered at each time as being produced in one of these states and for each phoneme state transitions are defined according to a Bakis model. Emission probabilities in each state are thus modelized as a non linear auto-regressive process of fixed order driven by a white noise whose parameters depend on the current state :

$$x_t = F_p(x(c_t), \Theta_p) + n_{p,t}$$

where F_p is the function computed by the p^{th} predictor with parameters Θ_p, $n_{p,t}$ is the prediction error in state p at time t which is assumed to be an independent and identically distributed (iid) random variable with probability density function $P(n_{p,t})$, and c_t is the prediction context at time t. If we assume that $n_{p,t}$ is gaussian with zero mean, and covariance matrix Σ_p, we can write:

0-7803-2431-5/95 $4.00

$$P(n_{p,t}) = \frac{1}{(2\pi)^{\frac{d}{2}} . |\Sigma_p|^{\frac{1}{2}}} . \exp(-\frac{1}{2} n_{p,t}^T \Sigma_p^{-1} n_{p,t}) \quad (1)$$

$$= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_p|^{\frac{1}{2}}} . \exp(-\frac{1}{2} (x_t - \hat{x}_p(c_t))^T \Sigma_p^{-1} (x_t - \hat{x}_p(c_t)))$$

where d is the dimension of the input vector, and $\hat{x}_p(c_t) = F_p(x(c_t), \Theta_p)$. The likelihood of an acoustic vector sequence (X_1^T) along a particular state sequence (s_1^T) is computed as [5]:

$$P(X_1^T, s_1^T) = \prod_{t=1}^{T} P(x_t - F_{p_t}(x(c_t), \Theta_{p_t})) P(s_t | s_{t-1})$$

For training the NPS, one starts from an initial value of the predictors parameters and perform an iterative algorithm which alternates the following two steps:

1. computation of model scores and segmentation.
2. optimization of NN parameters.

Let $L = -\log P(X_1^T, s_1^T)$ and θ_p be a parameter from predictor p. Optimization is performed according to a simple gradient algorithm:

$$\theta_p = \theta_p - \varepsilon \frac{\partial L}{\partial \theta_p}$$

The gradient can be decomposed as:

$$\frac{\partial L}{\partial \theta_p} = \sum_{j=1}^{d} \frac{\partial L}{\partial \hat{x}_{p,j}} \frac{\partial \hat{x}_{p,j}}{\partial \theta_p} \quad (2)$$

where j indexes the outputs of the p^{th} predictor. Let $t_1 ... t_{p_n}$ be the time sequence corresponding to the occurrences of state p, or equivalently predictor p, in the sequence s_1^T. It follows :

$$\frac{\partial L}{\partial \hat{x}_{p,j}} = - \sum_{i=1}^{p_n} \sum_{k=1}^{d} \sigma_{kj} (x_{t_i,k} - \hat{x}_{p,k}(c_{t_i})) \quad (3)$$

where σ_{kj} is the $(k*i)^{th}$ element of Σ_p^{-1} and d is the input dimension.

The second term of the right hand side of (2) can be computed through off line back propagation. (3) thus allows to compute the gradient of L with respect to any parameter θ of a predictive network.

3. DISCRIMINATION AT THE FRAME LEVEL

3.1. Implementations

We have proposed [6, 7] discriminative neural predictive systems (DNPS) for acoustic-phonetic decoding in CSR. Simple discriminative training at the frame level [6] raised the performances of NPS up to the state of the art (74.9% of phonetic accuracy). This work and others proved the feasibility of the predictive NN approach for CSR. As an example, recognition results for each of the phonetic classes are given in Table 1, for the neural predictive system (NPS) and our best local discriminant system (DNPS). Experiments have been performed on 88 TIMIT database [8]. The DNPS was more accurate than the NPS, with improvements ranging from 3.6% for the glides to 11.6% for the nasals. The Improvements are more significant for stops and fricatives. Most errors are performed on vowels essentially due to their duration.

phonetic classes	% NPS	% DNPS	
Stops	56.2	67.8	+ 11.6
Fricatives	63.1	74.6	+ 11.5
Nasals	73.5	79.3	+ 5.8
Vowels	48.1	57.2	+ 9.1
Glides	79.4	83.0	+ 3.6
Silences	95.9	96.1	+ 0.2

Table 1. Performances (% correct) on phone recognition for the six phone classes in the NPS and the DNPS. A bigram phone-language model has been used and insertions were not considered as errors .

3.2. Probabilistic interpretation of frame discriminative training

In its simplest form, discrimination has been implemented at the frame level by increasing the output of the correct model while decreasing the output of all other models. This simple discriminative criterion can be derived as shown below through the maximization of mutual information evaluated over the sequence of correct models. Let I be this mutual information:

$$I = \sum_{t=1}^{T} \log \frac{P(x_t, p_{i_t})}{P(x_t)P(p_{i_t})} = \sum_{t=1}^{T} \log \frac{P(x_t \mid p_{i_t})}{\sum_{i=1}^{Q} P(x_t \mid p_i)\, P(p_i)} \quad (4)$$

where x is a sequence of T speech frame vectors, $x_1,..,x_T$, Q the number of predictors, and p_{i_t} the correct predictor for frame x_t.
The priors $P(p_i)$ can be estimated as long-term statistics from data, for simplicity, we will assume that all states are equiprobable and that $n \sim \mathcal{N}(0,I)$. Equation (1) then writes :

$$P(x_t \mid p_{i_t}) \alpha \exp(-\frac{1}{2} D^2_{i_t})$$

where $D^2_{i_t}$ is the euclidean distance between desired and computed outputs for the correct predictor at time t, and (4) becomes :

$$I = \sum_{t=1}^{T} \log \frac{e^{-D^2_{i_t}}}{\sum_{i=1}^{N} e^{-D^2_{i_t}}} \quad (5)$$

Training the system to maximize (5) over all frame sequences obviously implements frame discrimination or frame-MMI training.

However this form of discrimination is computationally heavy. It has been found much more efficient, both for training time and accuracy, to perform discrimination between the correct model and selected nearest competitors at each time instead of all competitors [6].

4. DISCRIMINATION AT THE PHONE LEVEL

4.1 A simple approach

Discrimination at the frame level is clearly not optimal since it takes into account only local information. Discrimination should be performed at the sequence level, by considering competing phoneme models which correspond to the most likely paths for Viterbi scoring. This would allow to concentrate on models which actually compete during the decision process, which is not the case with frame discrimination. In [7] we have presented a simple strategy for doing this. During training, we considered the two sequences of phonetic models corresponding respectively to the real labelling of the sentence and to the system decision. Discrimination was performed only on models which did not agree along the two sequences. This simple technique did not led to significative improvements while being much heavier computationally than simple frame discrimination.

4.2. N-Best based discrimination

We have then implemented competition between models corresponding to the N most likely paths computed by an N-Best algorithm. This should provide a better approximation of the classification risk for sequences. We have implemented different criteria and undertaken a series of experiments for measuring the performances of different N-Best algorithms for this paradigm. Finally, we have used a simple version of the N-Best method initially proposed in [9, 10]. This implementation is computationally efficient for our problem, it is briefly described below.

Algorithm

Let:
c(t,i,j) be the cost of moving from state i at time t-1 to state j at time t.
$C_k(t,s)=s(0)....s(t)$ be the k^{th} best path from the state $s_d=s(0)$ to state $s=s(t)$ and $D_k(t,s)$ its accumulated cost.
Ph(C) the phone sequence associated to a sequence C.
$\otimes$ the concatenation operator.
For a sentence with T acoustic frames, we obtained the N-Best phone sequences as follows:

Initialisation :

$D_k(0,s) = \begin{cases} 0 & : \text{if } k = 1 \text{ and } s = s_d \\ \text{infinity} & : \text{otherwise} \end{cases}$

$C_k(0,s) = \begin{cases} s_d & : \text{if } k = 1 \text{ and } s = s_d \\ \text{empty} & : \text{otherwise} \end{cases}$

Recursion over t=1,...,T:
For k = 1,..., N
$D_k(t,s) = \min_{k',s'} (D_{k'}(t-1,s') + c(t,s',s))$ such that $Ph(C_{k'}(t-1, s') \otimes s) \neq Ph(C_m(t, s))$ for all m<k)
The minimum is thus taken over those state sequences which are not yet in the list of the (k-1) best phoneme sequences saved at the level of node t.
$C_k(t,s) = C_{k'}(t-1, s') \otimes s$, where k' and s' minimize $D_k(t,s)$.

This algorithm ensures that $Ph(C_2(T,s)) \neq Ph(C_1(T,s))$. In our implementation we have forced this difference to be larger than a given threshold.
In order to reduce computing time, memory requirements and to prune unlikely paths, we have used a beam-search strategy. Only the paths whose cost remain in a given "beam" around that of the best path are kept as candidates [11]. Those that fall outside the beam are pruned. In the experiments on TIMIT, we have considered only two best paths (Figure 1). The models which belong to the first path will be considered as correct and trained so as to increase their accuracy, while those which belong to the second path and not to the first will be considered incorrect and updated so as to increase their error. Sequence discrimination slightly increases the performances compared to frame discrimination (Table 1).

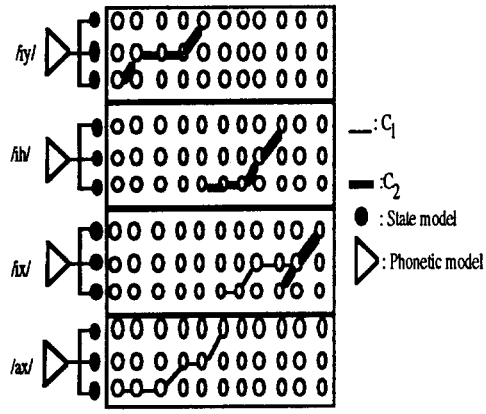

Figure 1: The phonetic sequence of C_1 is different from C2. Their differences are based on phonetic model units and not on state models.

We have also incorporated into the model learned transitions costs ($P(s_i | s_{i-1})$). The system is trained to learn intra-model state transitions by a simple counting method. This also increased the performances (Table 1) at a little extra cost.

system	2-Best	2-Best with transition probabilities
Phone Substitutions(%)	14,2	13,4
Phone Insertions(%)	4,6	4,3
Phone Deletions (%)	5,2	4,9
Phone Correct(%)	81,6	81,7
Phone accuracy (%)	76,0	77,4

Table 1: Performances of our system with 2-Best discrimination criterion and transition probabilities on TIMIT database.

5. CONCLUSION

After describing a general formalism for training neural predictive systems, we have presented several implementations for discriminative training. Global criteria measured on whole sequences allow to reach state of the art performances. Local discrimination is somewhat below but behaves remarkably well at a low computational cost.

6. REFERENCES

[1] Tebelskis J., Waibel A., Petek B., Schmidbauer O.:" Continuous speech recognition using linked predictive neural networks", ICASSP 91, pp 61-64, 1991.

[2] Iso K., Watanabe T.: " Large vocabulary speech recognition using neural prediction model", ICASSP 91, pp 57-60, 1991.

[3] Mellouk A., Gallinari P.:"A discriminative neural prediction system for speech recognition", ICASSP 93, ppII553-536, 1993.

[4] Makhoul J.: "Linear prediction: a tutorial review", Proccedings of IEEE, Vol. 63, N° 4, Apr 75.

[5] Mellouk A.:" A neural predictive system for continuous speech recognition" Phd Thesis, University of Paris-Sud, 1994.

[6] Mellouk A., Gallinari P.: "Continuous speech recognition systems", ICANN 93, pp 383-388, 1993.

[7] Mellouk A., Gallinari P.: "Discriminative training for improved neural prediction system", ICASSP 94, pp I233-I236, 1994.

[8] Grafolo J.S. 1988: "Getting Started with the DARPA TIMIT CD-ROM: an Acoustic Phonetic Continuous Speech Database", NIST, Gaithersburgh, Md.

[9] Schwartz R., Chow Y.L.:" The N-Best algorithm: An efficient and exact procedure for finding the N most likely hypotheses" In ICASSP 90, pp 81-84, 1990.

[10] Steinbiss V.:" Sentence-Hypotheses generation in continuous speech recognition system" In EuroSpeech, pp 51-54, 1989.

[11] Lowerre B.T., Reddy D.R.:" The HARPY speech understanding system" In W. A. Lea, ed., Trends in Speech Recognition. Englewood Cliffs, N.J.: Prentice Hall, 1980.