PREDICTIVE NEURAL NETWORKS IN TEXT INDEPENDENT SPEAKER VERIFICATION: AN EVALUATION ON THE SIVA DATABASE

A. Paoloni, S. Ragazzini, G. Ravaioli

Fondazione Ugo Bordoni
Via Baldassarre Castiglione 59, 00142 Roma

ABSTRACT

In this paper we propose a system which combines the use of predictive neural networks and the statistical approach in the task of text-independent speaker verification through the telephone line. The system is composed by a predictive neural network for every reference speaker, which is trained with the back-propagation algorithm and the maximum likelihood criterion, in order to obtain the highest probability when the input to the network belongs to the reference speaker.

We also consider a global network trained on the whole training set whose likelihood gives a measure of the predictability of a given input with the aim to eliminate the strong dependence of the score from the particular input considered.

The evaluation of the system is carried out on a subset of the Italian telephonic database SIVA, purposely collected for the considered task.

1. INTRODUCTION

Speaker verification consists in to determine if a voice sample truly belongs to the person he claims to be. A generic speaker verification system accepts the speaker's identification code and his voice sample, makes a comparison of the distinctive features of this speaker with that of the reference speaker corresponding to the given identification code. If the result of this comparison is below a fixed threshold the unknown speaker's identity is confirmed as whom he claims to be.

A speaker verification system can make two different types of error i.e. to deny the access to the true speaker (false rejection) or to allow the impostor to be accepted as the true speaker (false acceptation). Reducing the acceptation threshold grows the number of false rejection (FR) while the number of false acceptation (FA) decreases; the intersection of these two error functions, is known in technical literature as Equal Error Rate (EER), and corresponds to make a compromise with the need of a high security level and a low number of trial to access for the true speaker.

A speaker verification systems can works in two different manners, in text-dependent modality if the speaker is required to utter a keyword or a password, or in text-independent one if the identification process is not connected to a particular word or phrase.

In the following is described a system based on the application of predictive neural networks (PNN) [1,2,3], one for each reference speaker, trained with the back-propagation algorithm and the maximum likelihood criterion, in order to obtain the highest probability when the input to the network belongs to the reference speaker.

We also consider a normalisation network trained on the whole reference set whose likelihood gives a measure of the predictability of a given input. With a view to doing a normalisation to eliminate the strong dependence of the likelihood from the particular input considered, we subtract from each reference network log likelihood the global network one. In fact, when a speech segment contains an unpredictable factor, the prediction error grows for both the each reference network and the global one; considering the ratio between these two errors we obtain an elimination of the influence of the unpredictable factor.

2. PREDICTIVE NEURAL NETWORKS

The predictive neural network is a non-linear prediction model based on Multi Layer Perceptrons. PNN non-linearly predicts a parameter vector from several context frames. The main advantage of this approach is that as a PNN is trained for each reference speaker, to add or to remove a reference speaker we only need to train or to delete the corresponding network.

When a MLP [4] is used as a predictor of order p, it is trained to approximate in the best way the target $\mathbf{t}^*=\mathbf{y}_{t+1}$ from p parameter vectors that are the input to the network $\mathbf{x}^*=\{\mathbf{y}_t, \mathbf{y}_{t-1}, \cdots, \mathbf{y}_{t-p}\}$; as depicted in Fig.1 the neural network realizes the input-output mapping:

$$(2.1)\quad \hat{\mathbf{y}}_{t+1} = F_\omega(\mathbf{x}^*) = F_\omega(\mathbf{y}_t, \mathbf{y}_{t-1} \cdots, \mathbf{y}_{t-p})$$

where ω means the set of the parameters that characterize the network, i.e. the connection weights and the value of the parameter of the sigmoidal function used as non-linearity.

2.1. Statistical interpretation of PNN

Usually PNNs are trained minimizing the distortion between the target and its prediction sequences:

$$(2.2)\quad E = \sum_{t=p}^{T} \left\| \mathbf{y}_{t+1} - F_\omega(\mathbf{y}_t, \mathbf{y}_{t-1} \cdots, \mathbf{y}_{t-p}) \right\|^2$$

It is possible to model speech production as a non linear autoregressive process and to give a statistical interpretation to the model [5,6]:

$$(2.3)\ y_{t+1} = \hat{y}_{t+1} + \varepsilon_t = F_\omega(y_t, y_{t-1}\cdots, y_{t-p}) + \varepsilon_t$$

where F_ω is the function computed by the model of the talker, ε_t is a noise whose realizations are considered independent.

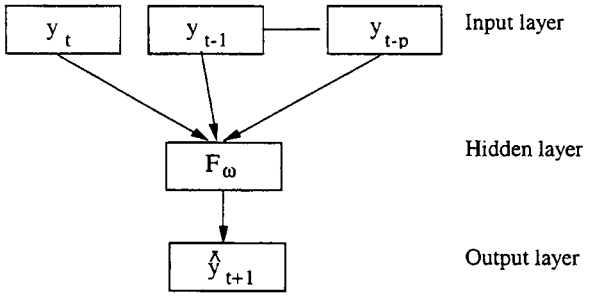

Figure 1: Realization of a p order predictor by mean of a three layer neural network that approximates the transfer function F_ω.

We suppose that all speakers have the same a priori probability and that ε_t is gaussian for each talker, with mean μ and covariance matrix Σ

$$(2.4)\ p(\varepsilon_t) = \frac{1}{\sqrt{(2\pi)^M \det\Sigma}} \exp\left\{-\frac{(\varepsilon_t-\mu)\Sigma^{-1}(\varepsilon_t-\mu)^T}{2}\right\}$$

The conditional likelihood that a given speech segment lasting L frames belongs to i-th reference speaker, whose parameter network are Γ, is expressed by :

$$(2.5)\ P(\mathbf{Y}^L/\Gamma) = \prod_{t=1}^{L} p(\varepsilon_t) = \prod_{t=1}^{L} p(y_{t+1}/y_t\cdots y_{t-p},\Gamma)$$

Training the neural network through a distance error criterion, taking into account the statistics of the prediction error, leads to a statistical interpretation as maximum likelihood training procedure. We chose to work in the logarithmic domain in order to simplify likelihood computation

$$(2.6)\ V_i = V(\mathbf{Y}^L/\Gamma) = -\log_e P(\mathbf{Y}^L/\Gamma)$$

In fig.2 is depicted the verification scheme for a reference speaker.

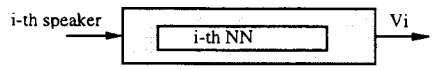

Figure 2: Verification of i-th speaker without error normalisation.

If the likelihood V_i is below the score threshold, the speaker is accepted as the i-th speaker.

2.2. Likelihood normalisation

Although the likelihood is less sensible than the prediction error to unlearned factors that can be an unusual noise, or some difference of the speaker's voice between the training and testing data, as in the case of a noisy channel.

This phenomenon is well known in technical literature [2,6], and it is resolved considering a global PNN trained for multiple categories, i.e. for all the reference speakers in order to obtain a predictability measure. In fact when a speech segment contains an unpredictable factor, the prediction error grows in the same manner both for the single speaker PNN (ε_{ti}) and the global network (ε_g), considering

$$(2.7)\ \varepsilon'_{ti} = \frac{\varepsilon_{ti}}{\varepsilon_g}$$

we obtain the normalised prediction error which exhibits a lower dependence from the unpredictable factor. Considering the log likelihood formula (2.6) becomes

$$(2.8)\ V'_i = V(\mathbf{Y}^L/\Gamma) - V(\mathbf{Y}^L/\Gamma_g) = V_i - V_g$$

where Γ_g are the parameters of the global network.

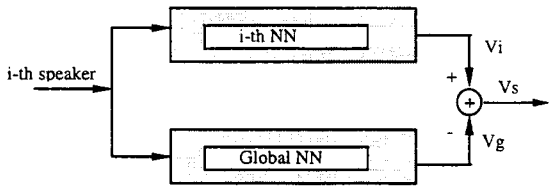

Figure 3: Verification of i-th speaker with error normalisation. The normalisation is performed by the global network, trained using all the reference speakers.

In practice the speaker to be verified is tested both on the network of the speaker he claims to be and on the global network. If the difference between the single network and the normalisation network is lower than a fixed threshold, the identity of the speaker is confirmed. With this scheme if it is necessary to add a new reference speaker, we have to train again the global network, but as it is demonstrated in [2], if the normalisation network was trained on a large data set of many reference speakers, the performances of the system are only slight worse if we don't re-train the global network.

3. EXPERIMENT DESCRIPTION

The experiments described in the following were carried out on a subset of the Italian telephonic database SIVA. The SIVA database has been collected over the Public Switched Telephone

Network (PSTN), using several types of telephonic handsets [7]. It contains 18 repetitions of 20 male speakers.

Each session contains a list of isolated words, a dialogue and a read passage, for a total of about 180 seconds. Only the read passage is used in this work.

The speech was sampled at 8 kHz in PCM μ-law 8 bits format. the signals have been preenphatised with a factor of 0.95, and frames have been weighted using the Hamming window.

The FFT analysis was done on a window of 64 ms, sliding of 10 ms. 13 Mel-scaled coefficients were computed from the FFT spectrum and used as input to the PNN.

On the cepstral data was performed a mean subtraction in order to improve the prediction capability of the PNNs.

The PNN was trained to perform a prediction of order three, that is to say to predict a frame from the preceding three.

The neural network was a three layer feedforward network with 39 input nodes, 11 hidden nodes and 13 output nodes. The network for every speaker was trained with the backpropagation algorithm assuming the prediction error has an initial mean zero and its variance is set to unity.

A speaker set of 12 speakers was chosen as reference, and 8 speakers as impostors.

Each speaker uttered 18 repetitions of the phonetically balanced passage, a repetition has a mean duration of 60 seconds. For the reference speakers 9 repetitions were randomly selected for the training phase. From these 9 repetitions we choose some frames in order to obtain an independence from the text of the passage and to take into account the different conditions of the speaker and of the telephone during the time.

We considered different lengths for the training data formed by several frames taken from these nine utterances as reported in the following section.

One network was trained for each reference speaker, and the global network, which has the same structure and dimension as the single speaker's network, was trained on the data of all the reference speakers.

The verification test was performed on 18 repetitions for the 8 impostors and on 9 repetitions out of the training set for the 12 reference speakers.

To evaluate the performances of the system we consider a pooled distribution of all the true customer scores versus the impostor scores that shows the degree of separation of the two classes with a single threshold. In this way we found a pooled Equal Error Rate (EER) that gives us a measure of the goodness of the system. However a single threshold is not a good threshold for every speaker, it is possible to chose an individual threshold and to calculate EER' that is the mean of the individual EER. In the following section will be shown the results both for a pooled and individual threshold.

3.1. The verification and the likelihood normalisation

In this section we first analyse the capability of the error normalisation technique.

We trained a PNN for each reference speaker with different lengths of the training set with the scheme shown in fig.2, and evaluated the EER.

The training data were chosen from different repetitions of the reference speaker that is about 20 seconds of each repetition for a training set lasting 3 minutes.

Training duration	3 min.	2 min.	1 min.
EER	26.5 %	28.7 %	32.1 %

Tab. 1: Equal error rate for three training set lengths using a PNN for each reference speaker. The test was carried out on 30 seconds of speech.

On the same training set was trained the normalisation network, and the likelihood obtained with the individual networks were weighted with the global network scores as in formula (2.8) with the scheme depicted in fig.3. By comparing tab.1 and tab.2 we can note a considerable improvement of the system performances. For a training set lasting 3 minutes the EER decreases from 26.5% to 3.7 %.

Training duration	3 min.	2 min.	1 min.
EER	3.7 %	8.7 %	14.8 %

Tab. 2: Equal error rate for some training set length using the normalisation network. The test is carried out on 30 seconds of speech.

We chose to make the learning of the PNNs on a training set lasting 3 minutes, and we evaluated the performances of the system on different lengths of the test set. The results are shown in tab.3 for a pooled threshold score.

While the length of the training set isn't a critical problem, as the training of the PNNs is made off line, the length of the test set is a real problem, as we need to make a compromise with good performance and the customer's speech duration in the verification procedure.

Test duration	20 sec.	10 sec.	5 sec.
EER	4.5 %	7.5 %	9.6 %

Tab. 3: Equal error rate for three test set lengths using the normalisation network. The training was carried out on 3 minutes of speech.

From the analysis of the distribution of FA and FR of the reference speakers we can see very different distributions for the various speakers. Fig.4 shows the FA and FR distributions for the threshold that gives the EER, we can see that the speaker FR has

0% FA and 0% of FR, and other speakers as DA and LA have only FA while BE has only a high FA. This circumstance lead us to consider an individual threshold in order to improve the performance of the system.

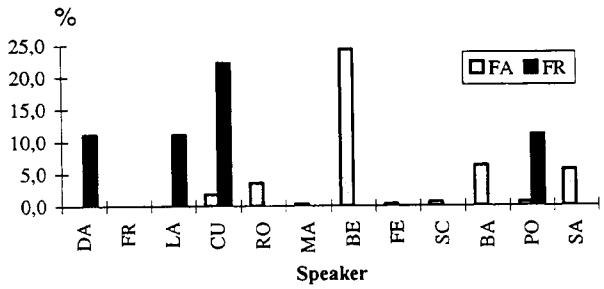

Fig.4: Distribution of False Acceptance and False Rejection for the threshold giving the EER. The training was carried out on 3 minutes and the test on 20 seconds of speech.

Considering an individual threshold we computed the EER' that is the mean of the individual EER depicted in fig.5. We obtain an high improvement in the performance of the system as is shown in tab.4.

Test duration	20 sec.	10 sec.	5 sec.
EER'	0.55 %	1.6 %	5.48 %

Tab. 4: Equal error rate for some test set length using the normalisation network. The training was carried out on 3 minutes of speech. In this case was used a threshold for each reference speaker.

Analysing the individual EER we can note that 5 speakers have an EER equal to 0%, and only 2 speakers have an EER equal to 2.4 , the resulting mean Equal Error Rate EER' is equal to 0.55%.

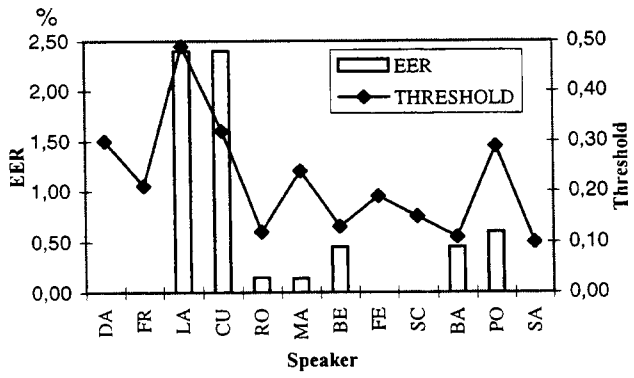

Fig.5: Distribution of the individual EER and corresponding threshold value for the reference speakers. The training was carried out on 3 minutes and the test on 20 seconds of speech.

The mean value of the individual threshold scores is very near to the pooled threshold score.

4. CONCLUSIONS

In this work we have presented a series of experiments considering a statistical interpretation of PNN's prediction error in the text-independent speaker verification on an Italian telephonic database collected over real telephonic network.

We have considered a likelihood normalisation technique that highly improve the verification capabilities of a pool of Predictive Neural Networks.

Considering an individual threshold score we reached a mean Equal Error Rate equal to 0,55 %.

Further improvements can be achieved considering multiple states for each speaker, that is to consider multiple networks for each reference speaker and an ergotic model for each speaker.

5. REFERENCES

1. H. Hattori, "Text-Independent Speaker verification Using Neural Networks", *Proceedings ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994, pp. 103- 107*

2. H. Hattori, "Text-Independent Speaker recognition Using Neural Networks", *Proceedings ICASSP'92 vol.II pp.153-156*

3. Y.Benanni, P.Gallinari, "A connessionist approach for speaker identification", *Proceedings ICASSP 1991, pp. 385-388.*

4. R.Lippman, " An introduction to computing with neural nets", *IEEE ASSP Magazine, pp.4-22,Apr.1987*

5. B. Petek, A. Ferligoj, "Exploiting Prediction Error in a Predictive-Based Connectionist Speech Recognition System", *Proceedings ICASSP 93, vol. 2, pp. 267-270.*

6. S. Ragazzini, G. Ravaioli, " Reti neurali predittive nella verifica del parlatore ", *Internal Report FUB 5B03095.*

7. M. Falcone, U. Contino, "Acoustic Characterisation of Speech Databases: an Example for the Speaker verification", *Proceedings of the 'International Congress on Phonetic Science', Stockholm, 1995, pp.290-294.*