

Improving the Competitiveness of Discriminant Neural Networks in Speaker Verification

C. Vivaracho-Pascual

U. de Valladolid, Spain
Dep. de Informática
cevp@infor.uva.es

J. Ortega-Garcia

U. Politecnica de Madrid, Spain
Biometrics Research Lab.
jortega@diac.upm.es

L. Alonso-Romero

U. de Salamanca, Spain
Dep. Inf. y Autom.
lalonso@tejo.usal.es

Q. Moro-Sancho

U. de Valladolid, Spain
Dep. de Informática
isaac@infor.uva.es

Abstract

The Artificial Neural Network (ANN) Multilayer Perceptron (MLP) has shown good performance levels as discriminant system in text-independent Speaker Verification (SV) tasks, as shown in our work presented at Eurospeech 2001. In this paper, substantial improvements with regard to that reference architecture are described. Firstly, a new heuristic method for selecting the impostors in the ANN training process is presented, eliminating the random nature of the system behaviour introduced by the traditional random selection. The use of the proposed selection method, together with an improvement in the classification stage based on a selective use of the network outputs to calculate the final sample score, and an optimisation of the MLP learning coefficient, allow an improvement of over 35% with regard to our reference system, reaching a final EER of 13% over the NIST-AHUMADA database. These promising results show that MLP as discriminant system can be competitive with respect to GMM-based SV systems.

1. Introduction

The promising results achieved in a previous comparative work with GMMs [1], has led us to deepen in the use of discriminant ANNs in SV, in order to improve the system. Such improvements are now presented in this contribution.

One of the most important problems still to be solved is the discriminant use of MLPs, as this implies learning to discriminate between target and non-target speakers (the universe of speakers different from target), by using training samples of both. This non-target class will be denoted as *Impostor Speakers for Training* (IST).

Regarding target training samples, the more there are the better they will be represented, thus all possible examples are used. However, the IST selection represents a problem, as the amount of candidates is, normally, very high (an example can be seen in section 2). Training with that lack of proportion between samples of target and non-target speakers leads the ANN to overlearn this second class, leading it to misreject. Then, a subset from the pool of candidates must be chosen. In order to achieve an optimum ANN learning, this subset must be optimum in content (as the most representative subset within those included in the pool of candidates must be determined) and in size (too few would be non-representative and the ANN would tend to misaccept, while with too many the network would tend to misreject). Traditionally, random selection has been used to IST selection, but, as it will be shown, the achievement of competitive systems is problematic with this methodology.

This work has been partially supported by Spanish Ministry of Science and Technology under project TIC2000-1669-C04.

The lack of research into this subject, and its importance to adjust the system performance, has led us to focus on it. In section 6, a new heuristic method of IST selection is presented, improving the system performance and allowing IST size optimisation.

It must be noticed that the IST selection problem could be seen similar to cohort selection or universal background model creation in GMM/HMM-based [2] or Autoassociative Neural Networks (AANN)-based [3] systems score normalisation. But, in practice, as these classifiers work very differently from discriminant ANNs¹, the problem is also different. Then, none of the proposed cohort selection nor universal background model creation techniques can be applied in the IST selection.

The second major contribution of this proposal can be found in section 5, where is proposed a simple, but effective, selective use of the ANN outputs for the computation of the final score per sample, in order to improve the sample classification.

Together with the above proposed procedures, we will be concerned with an optimisation of the ANN learning coefficient, which means that the improvement in performance, with respect to that of the system in [1] (the Reference System (RS)), is over 35%, achieving EERs of 13%, as will be seen in Section 9. A comparison with other systems performance can be seen in Section 10.

2. Speech Corpus

The subset of speaker recognition-oriented AHUMADA database [4] included in the NIST 00 and 01 evaluations is used. It is a telephonic subcorpus in Spanish, only with male speakers and channel mismatch between training and testing segments. The development set consists of 122 speakers, each of them with speech segments of about 30 s of duration. These speakers are the representatives of the non-target class, then the IST are selected from this set.

Due to computational load requirements for each test, the experiments are only carried out on 50 of the 103 target speakers contained in the corpus, considering that this randomly selected subset provides representative results.

3. Parameter Extraction

Standard feature extraction is accomplished through 14 mel-frequency cepstral coefficients (MFCC) plus 14 Δ MFCC vectors; these are extracted from 32 ms frames, taken every 16 ms with Hamming windowing. Pre-emphasis factor of 0.97 is re-

¹GMM/HMM/AANN model the training characteristic vector distribution, being the output a $P(x/\lambda)$ estimation. Discriminant MLP classification is based on linear discriminant functions between the training vectors of the different classes; the output can be seen as $P(\lambda/x)$.

alized, and cepstral mean normalisation (CMN) is applied as channel compensation scheme.

4. Reference System

A 3-layer MLP per speaker is trained, with 28 neurons in the input layer, 32 in the hidden one and 1 in the output. In order to reach a correct performance in the training algorithm, the values of each of the components x_{ij} of the input vector x_i are bounded. The alternative which has shown the best performance [1] is the so-called *vector normalisation*, that is:

$$\hat{x}_{ij} = \frac{x_{ij}}{\max_j(|x_{ij}|)} \quad (1)$$

The number of the *non-target* class vectors needed for an optimum MLP training is much higher than those of the *target* class. In order not to decrease the performance, both sets should be equal. As before, the solution that produces best results [1] is the *User Vector Repetition*, which consists in repeating each vector of the target, until both sets are equal. Error Backpropagation is used for training, with learning coefficient $\alpha = 0.1$, and momentum 0.95. The desired outputs for target speaker vectors is 1.0 and for the IST, 0.0.

5. Classification Improvement

Given a sample X with M vectors, $X = \{x_1, x_2, \dots, x_M\}$, the final score of the system will be calculated as:

$$S(X) = \frac{1}{M} \sum_{i=1}^M \log(P(\lambda_c/x_i)) \quad (2)$$

Where $P(\lambda_c/x_i)$ is the output of the user c MLP for input vector x_i .

This first proposal in order to improve the system is based on the idea of locating and eliminating the damaged parts from the speech, as proposed in the *Missing Feature Theory* [5]. Following this approach, those vectors whose output value is intermediate, i.e., not clearly identified with the target speaker or the impostor, shall be called the “noisy parts”. If we make both assumptions:

- Target samples produce MLP outputs usually high
- Impostor samples produce low outputs

We can eliminate these less-significant intermediate values for calculating the final score per sample, achieving, if not a direct improvement in the classification, at least an easier calculation of the decision threshold. (0.2,0.8) is chosen to be the interval of output values not considered. Applying the indicated proposal, the final result per sample will be:

$$S_R(X) = \frac{1}{M_R} \sum_{i=1}^{M_R} \log(P(\lambda_c/x_i)) \quad \forall x_i/P(\lambda_c/x_i) \notin (0.2, 0.8) \quad (3)$$

Where M_R is the number of vectors x_i that verify the rule in eq. (3), a rule that will be called from now on R262. In Fig. 1(a), the improvement produced by this procedure in terms of system performance is shown.

6. Heuristic Choice Proposal for the IST

Fig. 1(b) shows the main problems derived from random selection of IST: *i*) dependence of the result with respect to the selection made, and *ii*) impossibility of *a priori* determination of an optimum size for this set. A heuristic choice of the IST that should improve these limitations must comply with the two following conditions:

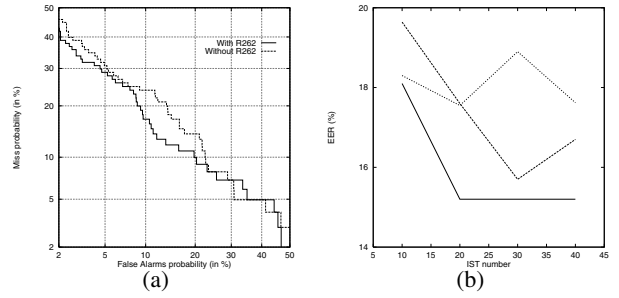


Figure 1: (a) Example of improvement with R262. (b) Performance evolution for 3 different IST random selections.

- *Robustness* with respect to the development set used
- *Effectiveness* compared with random choice results

The idea is to choose from among the IST candidates, typically the development set, the most discriminant subset, being the classifier itself, during the learning stage, who performs this selection. The proposed selection technique is based on the idea of *Incremental Learning* but, in this case, instead of modifying the network architecture during training, modifications affect to the composition of the training set, to be more precise, that corresponding to the set of the IST. The proposed technique can be formulated through the following steps:

- The size of both the IST, $MaxIST$, and the number of periods of network training, $MaxEp$, is selected.
- The network is trained using the target training samples and those of one speaker of the development set, over $MaxEp$ periods.
- $S(X)$ is obtained from the ANN for each of the remaining speakers of the development set, and the N with the highest values are chosen as the most similar to the target. If the network were used for classification, these N speakers would be those with the highest probability of producing false acceptances; though, due to their discriminant capability, these are selected to train the network in the next phase.
- The N speakers chosen in step c) are included into the IST set and the network is trained once again.
- Steps c) and d) are repeated until $MaxIST$ is reached.

In order to avoid random initialisation of IST, even the first IST (step b), is also heuristically selected. The closest speaker of the development set to the target should be chosen, so an AANN [6] will be employed, due to its capacity to model the training vectors distribution, here, the target training ones. Once training is complete, the development set speaker whose AANN output is closest to that obtained by the Target training samples will be used in step b).

Once the above process, referred from now on to as *Non-Target Incremental Learning* (NTIL), is completed and the MLP is ready for testing, the development set speakers not used in training are used to estimate the mean and variance of the impostor scores’ distribution in order to apply z-norm [2].

7. R262 and NTIL Performance Tests

The experiments carried out analyse the behaviour of the system with respect to the size of the IST set (several IST numbers were tested), the epochs of training and the value of N in the NTIL

technique. $N = 1$ and $N = 5$ values have been tested in order to compare results when the IST set is increased in one or in various (five) at a time.

Also, in order to compare the effectiveness of the NTIL proposal, 5 experiments were carried out with different random choices of the IST, namely A11 to A15. Due to the enormous quantity of results obtained for each of the tests above mentioned, only the most relevant will be summarised here.

7.1. Reference system results

Firstly, the RS results for the NIST - AHUMADA corpus are shown. Due to the inconclusive results obtained, not all the experiments initially proposed were completed.

Table I shows system performance, evaluated averaging the EER obtained in the training epochs 15 to 20 (best performance interval); if necessary, the epoch of ending the learning phase would be chosen from this interval. The results for different number of IST is shown.

	A11	A12	A13	A14	A15	A1
11 IST	23.4%	19.4%	22.6%	21.3%	22.0%	21.7%
16 IST	20.1%	22.0%	22.3%	22.1%	19.7%	21.2%
21 IST	20.2%	23.4%	-	-	-	21.8%

Table I: RS EERs for NIST-AHUMADA corpus.

7.2. Modification of the learning coefficient

Although in [1] the RS procedure seemed to provide good results, those shown in Table I for the NIST-AHUMADA are not, so modifications had to be introduced. Previous works showed some improvements with a decrease in the learning coefficient, assuming the cost of an increase in time for the learning phase of the network. Following this idea, the new value of α was fixed to 0.01.

Analysing the evolution of the system performance with respect to the training epoch, training the network for 150 epochs is chosen as the ending criteria for the learning stage, being the criteria applied in the remaining tests. Table II shows the results obtained, with an improvement in EERs of, on average, 15.3%.

	A11	A12	A13	A14	A15	A1
11 IST	19.0%	18.7%	18.3%	19.0%	17.2%	18.4%
16 IST	17.7%	19.0%	17.8%	18.5%	17.8%	18.1%
21 IST	19.0%	19.0%	19.6%	17.0%	17.0%	18.3%

Table II: EERs with $\alpha = 0.01$.

7.3. With R262

The proposal for improving the classification (Section 5) is now included, obtaining new improvements, as Table III shows.

	A11	A12	A13	A14	A15	A1
11 IST	17.9%	17.0%	17.0%	16.3%	16.0%	16.8%
16 IST	17.0%	16.5%	18.0%	17.0%	16.7%	17.0%
21 IST	16.5%	17.1%	18.6%	17.0%	16.0%	17.0%

Table III: EERs with $\alpha = 0.01$ and R262.

7.4. With NTIL

The technique described in Section 6 is now applied to obtain the IST, improving the previous results, as shown in Table IV.

For $N = 5$ not only the average performance of the random selection is improved, but also that of the best random case. Besides this advantage, the computational load of the training process is smaller for $N = 5$ than for $N = 1$.

	$N = 1$	$N = 5$
11 IST	17.0%	16.0%
16 IST	16.6%	14.0%
21 IST	16.2%	15.0%

Table IV: EERs with $\alpha = 0.01$, R262 and NTIL. The best results are emphasised with bold face.

8. NTIL Robustness Test

Once the effectiveness of the proposed scheme is proved, the aim of this Section is to demonstrate also the robustness of the NTIL procedure with respect to the development set; therefore, a new development set is used, while maintaining the same training and testing sets already used in the previous tests. Samples from *Speaker Recognition V1.1* database of CSLU (OGI) will now be used: changing language to English, under realistic recording (including background noise) and with a greater variety of handsets than in the original set (AhumadaDev). To be precise, from 3 to 4 segments of 36 male speakers (in order to reach the 122 files of the original set) are used, with an average speech duration of 12 s.

8.1. Results

Six tests have been carried out with different random selections, and 2 with NTIL: for N values of 1 and 10 (given the size of files, it was considered that using 5, as previously, was too small a step). Table V shows the results for IST set sizes (size measured in number of vectors) approximately equal to those of the previous experiments.

	Without R262			With R262		
	A1	$N=1$	$N=10$	A1	$N=1$	$N=10$
21 IST	17.9	17.0	16.3	17.0	16.0	15.2
31 IST	17.1	15.2	15.9	16.9	16.0	17.0
41 IST	18.3	18.0	16.3	17.0	17.0	13.0
51 IST	19.3	-	16.0	18.5	-	13.3

Table V: EERs (in %) with CsluDev. The best results are emphasised with bold face.

NTIL robustness have been shown, as the results are still better than those obtained with AhumadaDev. The results also prove the advantages of increasing the IST set in various at a time ($N > 1$) in NTIL technique, and the improvements with R262 procedure application (except in the case 31 IST, the only exception in all the experiments carried out).

9. IST Size Optimisation

Having improved performance, there is still one problem to be solved in the choice of IST: to find the optimum set size, as NTIL allows this parameter to be defined.

Once fixed the degree of freedom of the classifier (network architecture), the aim is to find the number of vectors of the training set that maximizes the capability of the ANN for generalisation. Since the target speaker training set size is also fixed, the concrete aim is to find the optimum number of vectors of the IST set: if it is too small it will not be representative and

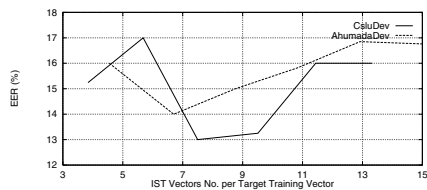


Figure 2: System performance evolution with IST size.

	NIST 99	NIST 00	NIST-AHUMADA
GMM	10.0%	11.0%	
AANN	14.4%	14.8%	
MLP			13.5%

Table VI: EERs with GMM-based and AANN-based SV systems applied on the evaluation databases of NIST 1999 and 2000, and with MLP applied on NIST-AHUMADA database.

the network will tend to misaccept; if it is too large, the network will overlearn and it will tend to misreject. Consequently, the optimum ratio between the number of vectors of the target speaker and the IST, in order to train the network, is searched.

Fig. 2 shows that system performance (with R262, $N = 5$ for AhumadaDev and $N = 10$ for CsluDev) with respect to the parameter to be determined is similar for both development sets, in spite of the great differences between them; and this reinforces the robustness of the NTIL selection. The best results (13-14% EER) are obtained, approximately, with 7 vectors of the IST for each target training vector, with an improvement of over 35% with respect to the RS; decreasing slightly with 9 and worsening thereafter.

It can also be derived from Fig. 2, if we compare these results with previous ones, that even for non-optimum IST sizes, the system performance is better (and, in the worst case, similar to) than the performance obtained using random selection.

10. Discriminant MLP competitiveness

Due to the impossibility to compare the showed results, with that achieved for other systems under the same experimental environment, since the NIST evaluation results are not public, the results shown in [6, 7] will be used as reference. The GMM-based system results shown in [6, 7] were taken from [8, 9]. These results can not be directly compared with those shown in previous sections, but can be used as reference to extract conclusions about the competitiveness of the discriminant MLP, as they are obtained under the same experimental conditions than ours: telephonic channel mismatch between training and testing samples and under NIST evaluation conditions, but with a different database.

In Table VI can be seen the results. GMM-based systems can be considered the state of the art, while AANNs is one of the most interesting alternatives to GMMs, as Yegnanarayana shows in [6]. Our system final result is that achieved in the optimum IST size, once shown that this parameter can be defined. The result shown in the table is the average of those obtained with the different development set used.

Using MLP the number of model parameters is as few as 961 (weights+bias), as opposed to 1847 used by AANN-based system and 9728 used by GMM-based system.

11. Conclusions

To reach a competitive system unpredictable behaviour must be eliminated. With ANN, one of the main causes of unpredictability is random selection of the IST. Applying the heuristic technique NTIL not only improves the performance of the system, but it also allows a parameter optimisation, independently of the development set used.

Although there is still work to be done, a key procedure has been proposed in improving the system, through the application of NTIL and R262 as a basic working architecture. From here onwards, new modifications and alternatives in the already tested procedure, and/or new techniques, will make the system surely improve. For example, initial tests using score normalisation has been accomplished with CsluDev set, though not used in discriminant networks, as this operation is supposed to be carried out in the learning stage, the preliminary results are highly promising as EER of just 11% has been achieved.

Finally, point out that the heuristic search technique NTIL can be easily generalized to any discriminant classifier.

12. ACKNOWLEDGMENT

We wish to acknowledge the CSLU (Oregon Graduate Institute) for providing us the Speaker Recognition V1.1 Database.

13. References

- [1] C. Vivaracho-Pascual, J. Ortega-Garcia, L. Alonso-Romero, and Q. Moro-Sancho, "A comparative study of MLP-based artificial neural networks in text-independent speaker verification against GMM-based systems," in *Proc. of Eurospeech01*. ISCA, 3-7 September 2001, vol. 3, pp. 1753-1756.
- [2] Roland Auckenthaler, Michael Carey, and Lloyd-Thomas Harvey, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.
- [3] S.P. Kishore and B. Yegnanarayana, "Speaker verification: Minimizing the channel effects using autoassociative neural networks models," in *Proc. IEEE ICASSP, Istanbul*, 6-9 June 2000, pp. 1101-1104.
- [4] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "An approach to forensic speaker verification using AHUMADA large speech corpus in spanish," *Speech Communication*, vol. 31, pp. 255-264, June 2000.
- [5] Andrzej Drygajlo and El-Maliki Mounir, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. ICASSP*, 1998, pp. 121-124.
- [6] B. Yegnanarayana and S.P. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459-469, April 2002.
- [7] S.P. Kishore, "Speaker verification using autoassociative neural networks," M.S. thesis, Indian Institute of Technology, University of Madras, <http://speech.cs.iitm.ernet.in>, 2000.
- [8] NIST (1999), "Speaker recognition workshop notebook," in *NIST 1999 Speaker Recognition Workshop*. NIST, 1999.
- [9] NIST (2000), "Speaker recognition workshop notebook," in *NIST 2000 Speaker Recognition Workshop*. NIST, 2000.