A Comparative Study of MLP-based Artificial Neural Networks in Text-Independent Speaker Verification against GMM-based Systems[1]

Carlos E. Vivaracho, Javier Ortega-Garcia, Luis Alonso, Quiliano I. Moro*

*Departamento de Informática
Universidad de Valladolid, Spain
cevp@infor.uva.es

Abstract

Text-independent speaker verification is an interesting task where the use of Gaussian Mixture Models is almost a must. Nevertheless, some preliminar encouraging results obtained in previous works using ANN in speaker verification have led us to consider to perform a direct comparison between these different methods. In this sense, this paper is only focused on the classification stage of both GMM-based and ANN-based speaker verification systems. Experiments are accomplish making use of the AHUMADA/GAUDI spanish speech database, specially oriented for speaker-recognition tasks as it contains multisession and multichannel data of about 500 speakers. Results confirm a better performance when using GMM-based system and microphonic speech but, on the other hand, when testing in specific conditions and with real telephone speech ANN outperforms GMM results.

1. Introduction

Different methodologies have been proposed in the classification stage of speaker recognition (SR) systems. Reviewing the state-of-the-art references, we can notice that Gaussian Mixture Models (GMMs) in text independent recongnition, and Hidden Markov Model (HMMs) in text dependent recognition, are the schemes more widely used. There are several comparative studies that support that tendency [1][2], but none of them related with Artificial Neural Networks (ANNs).

In the past decade, lot of works proposed the use of different ANNs paradigms in SR task [3,11], with successful performance in most of them. However, in the last years, the use of ANN in SR has decreased substantially. No objective reasons can be easily found, that explain that situation.

Our experience also confirms that encouraging results can be found when using ANNs[4,5]. In this sense, this work is focused in performing a comparative study trying to see whether ANNs still have chance in SR, or if the exposed recent tendency is verified. Specifically, we have used a Multilayer Perceptron (MLP) with backpropagation learning algorithm. There are different ways to use ANNs in SR [3]. We used the MLP as classifier: the output of the network is the a posteriori probability $P(\lambda/x)$, where λ is the speaker model, and x is the input vector.

Performance comparison between GMM and MLP in text independent speaker verification is shown in this contribution. Specifically, we reproduce some of the experiments performed by Ortega-Garcia *et al.* in [6] (reference system) using GMM, but this time using MLP-based ANNs. The only difference between the two works is in the classification stage, being the rest of the experimental environment the same.

The experimental environment will be exposed in section 2. In section 3 a complete description of the systems to compare will be found, with special emphasis regarding the MLP-based one. Results will be shown in section 4, and conclusions will be finally presented in section 5.

2. Experiment Description

2.1. Speech Corpus

Experiments have been accomplished by using GAUDI /AHUMADA speaker recogniton-oriented database [7], which consists in 100 male and 100 female speakers, multi-session, multi-channel, microphonic/telephonic huge database.

2.2. Parameter extraction

All microphonic speech has been down-sampled to 8 kHz (from the original sampling frequency of 16 kHz). 14 Mel-frequency cepstral coefficients (MFCC) plus 14 ΔMFCC have been used as feature vectors in all cases. Frames of 32 ms. taken every 16 ms. with Hamming windowing and pre-emphasis of 0.97 are used.

2.3. Experiments

9 experiments have been performed, 3 of them using microphonic speech (exps. 1, 2 and 3) and the rest using telephonic speech (exps. 4 to 9). A subset of 25 speakers from GAUDI/AHUMADA database are considered client speakers (users), and 25 other speakers operate as impostors. Tasks b (10 digit strings of 10 digit each, namely b01:b10) and c (10 fixed utterances, namely c01:c10) are used for training and/or testing the system. Specifically, c01:c5 are used for training, while c06:c10 and b01:b10 from user and one random utterance per impostor (from c:06:c10) are used to test the system. The experiments performed are the following:

- **Exp1.** Test utterances are recorded with the same microphone and in the same session with respect to speech material used in learning, being this case the most favourable situation.
- **Exp2.** Concentrates on inter-session variability, as the microphone used to train and test is the same.
- **Exp3.** The microphone and session to train and to test are different: microphone variability is added with respect to Exp2.

[1] This work was supported by Ministerio de Educación y Ciencia (MEC) grant TIC2000-1669-C0403.

- **Exp4.** T1 telephonic session (every speaker was calling from the same telephone, in a internal-routing call) is used to train, and T3 telephonic session (local call was made from a quiet room using 10 different standard handsets) to test.
- **Exp5.** T2 telephonic session (speakers made a local call from their own home telephone) was used to train, and T3 to test.
- **Exp6.** The number of train utterances is increased: T1 and T2 sessions are used now to train, and T3 to test.

In previous tests [6], Exps. 4 to 6 did not exhibit complete telephonic consistency. In order to verify telephonic consistency of the previous telephonic experiments, other experiments were proposed, making use of real telephonic speech from GAUDI/AHUMADA female subcorpus, namely T4, T5 and T6 sessions, all of them obtained in a real local-call acquisition process.

- **Exp7.** The system is trained with T4 and tested with T6.
- **Exp8.** In this case, T5 was used to train and T6 to test.
- **Exp9.** As in Exp6: T4 and T5 sessions are used to train and T6 to test.

3. Systems Description

3.1. GMM based system

Described in [6], it is the reference system. The distribution of feature vectors extracted from each user is modelled by a weighted *M* Gaussian mixture density:

$$p(x/\lambda_S) = \sum_{i=1}^{M} p_i^S b_i^S(x)$$

Tests with and without likelihood-domain normalization [8] were accomplished. As the density at point *X* (input sequence) for all speakers other than the true speaker is frequently dominated by density for the nearest reference speaker, nearest reference speaker normalization criterion was applied:

$$\log L(X) = \log p(X/\lambda_{S_c}) - \max_{S \in ref, S \neq S_c} \log p(X/\lambda_S)$$

Where λ_{Sc} means claimed speaker model. Balance between false rejection and false alarm error was required in order to calculate equal error rate (EER) for each speaker. Average EER through all speakers for each case was the final system score.

In spite of that tests with and without score normalization were accomplished, in the present work only score-normalization perspective is used, as results are clearly better.

4. MLP Based System

A different network per client speaker is trained, with samples of the user together with samples of other speakers that stand for the "rest of the world". This second set of speakers is the same that the so called "reference speakers" in the GMM based system: the another 24 user speakers.

The MLP has a three layer architecture: 28 neurons in the input layer, a variable number in the hidden one (performance with 4, 8, 32, 64, 96 and 128 neurons has been tested), and 1 neuron in the output layer, whose desired output is fixed to 1.0 and 0.0 when input vector belongs to client and impostor respectively.

The network output can be consider as the a posteriori probability $P(\lambda_S / x_i)$ that the input vector i from input sequence *X* belongs to the user *S* (let us identify the user S network as λ_S). Then, the probability that a input sequence *X* belongs to the speaker *S* will be:

$$P(\lambda_S / X) = \prod_{i=1}^{n} P(\lambda_S / x_i)$$

where *n* is the number of vectors in the input sequence *X*. The final value of that probability is very small; this is the reason why a log-score is used, being the final score O(*X*):

$$O(X) = \log P(\lambda_S / X) = \sum_{i=1}^{n} \log P(\lambda_S / x_i)$$

During training we stop after some learning iterations, and then we test the performance of the system (EER is evaluated). When the maximum iteration number is reached, the best performance weights values are kept. This procedure is repeated for different number of hidden layer neurons, and keep the best of all. This is the final measure of the performance of the system: the final EER that will be shown in the final section.

4.1. MLP training considerations

Due to the nature of the data, we had to face two practical problems in the network training.

First of all is the lack of proportion in MLP training sets: there is a ratio of 24:1 between the number of impostor and the number of user training vectors. Hence, the classifier tends to learn that "everything" must be rejected, and the system performance decreases.

Two different solutions have been tested:

- **Impostor Vectors Reduction (IVR)**. Used in [9], this solution is based on decreasing the amount of impostor vectors by clustering them, and using a representative of each cluster to train the network. The number of clusters was chosen to be, more or less, equal to the number of user training vectors. The clustering was performed by a self-organizing Kohonen map [10].
- **User Vectors Repetition (UVR).** The second solution tested consists in increasing the number of user vectors in the network training, repeating them as many times as training impostor speakers have been used. The goal is the same that in the other solution: to balance the number of vectors in the two training sets.

In order to verify the need to equalize the number of vectors, No Equalize, UVR and IVR options were tested in exp1. Table 1 shows the results. In the next section we will see the influence of the two proposed solutions.

No Equalizing	**UVR**	**IVR**
16.3 %	1.8 %	1.9 %

Table 1: EER in Exp1 without equalizing the size of the training vectors, UVR and IVR.

The second problem is referred to the normalization of the input data. Vector components have a wide range of values and it is necessary to normalize them in a range of [-1,1] in order to make the network work adequately. The following options have been tested:

- **Vector Normalization (VN).** Each vector (row) component is divided by the absolute value of the maximum one:

$$\hat{x}_{it} = \frac{x_{it}}{\max_{t}(|x_{it}|)}$$

where we are depicting the set of vectors *X* as a matrix, where the rows are the vectors. Then, let x_{it} be the component *t* of vector *i* extracted from input sequence *X*. $\hat{x}_{it}$ is the normalized component.

- **Column Normalization (CN).** Each column component is divided by the absolute value of the maximum one:

$$\hat{x}_{it} = \frac{x_{it}}{\max_{i}(|x_{it}|)}$$

- **Matrix Normalization (MN).** Each matrix component is divided by the absolute value of the maximum one:

$$\hat{x}_{it} = \frac{x_{it}}{\max_{it}(|x_{it}|)}$$

Other normalization schemes have also been tested, i.e. zero mean and unity variance *X* column vector normalization, but they did not outperformed the previous linear normalizations. Moreover, in fig. 1 we can see a comparison among VN, CN and MN, where it is plain that CN decreases its performance with respect to the other proposed schemes, being this the reason why only results obtained by means of VN and MN are considered from now on.

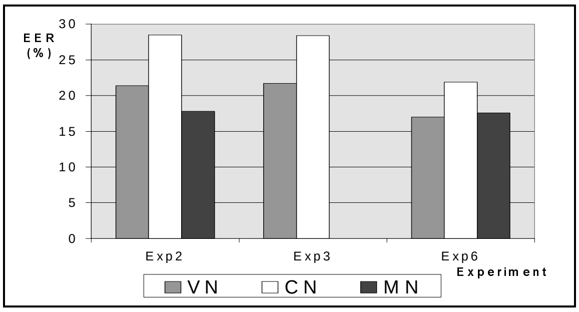

*Figure1. **System performance with VN, CN and MN.***

5. Results

As it was stated in section 3.1, the reference system was evaluated using EER. Tests with and without CMN (Cepstral Mean Normalization) channel compensation were performed in all of the experiments, except in exp1, because exp1 have no channel effects.

5.1. Microphonic speech

Following [9] we began the work using the IVR scheme to equalize the training sets size. In table2 we can see the first results. VN was used.

EER(%)	NONE		CMN	
	GMM	**MLP**	**GMM**	**MLP**
Exp1	0.2	1.9		
Exp2	12.0	17.9	10.6	21.4
Exp3	21.7	16.3	8.5	21.7

Table 2 GMM/MLP systems performance, with and without (NONE) CMN channel compensation. IVR and VN are used in MLP system.

It is obvious from table 2 that MLP performs worst, fundamentally when CMN is used. Besides, unexpectedly, CMN does not seem to compensate for channel effects when MLP is used. This was also observed in the first telephonic speech experiments. Trying to improve MLP performance, UVR and MN were tested. First, we performed same experiments trying to see tendencies. Fig. 2 and table 3 show the results of those experiments. Fig. 2 includes telephonic experiments. All of the experiments were performed with CMN channel compensation (except exp1), where GMM have demonstrate the best performance.

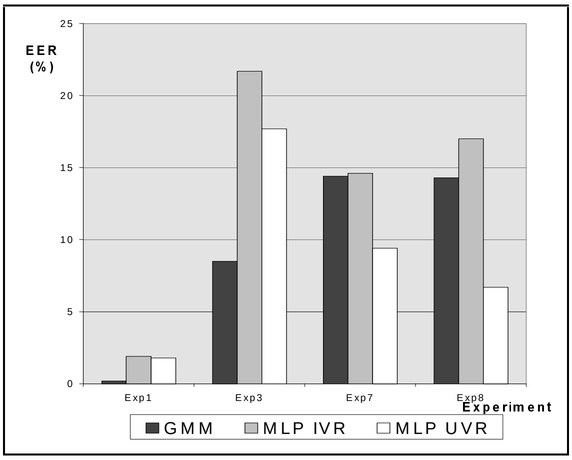

Figure 2. System performance with IVR and UVR**.** Input vector are VN

EER(%)	UVR		IVR	
	VN	**MN**	**VN**	**MN**
Exp3	17.7	19.1	21.7	25.6

Table 3 Results with VN and MN in exp3 with MLP based system. Data are CMN.

Regarding the results shown, it seems that GMM is outperforming MLP when microphonic speech is used. However, as we can see in fig. 2, MLP seems to outperform GMM when using telephonic speech. Hence, efforts are concentrated in order to verify this statement.

5.2. Telephonic speech

Previous results show clearly that UVR performs better than IVR, so these experiments make use only of that equalizing scheme. With regard to the normalization options (VN and MN), there were not, at this point, conclusive results; therefore, both normalization schemes are used to test.

Table 4 shows the results in exps. 4, 5 and 6, where telephonic sessions did not exhibit complete telephonic consistency.

EER(%)	NONE			CMN		
	GMM	**MLP VN**	**MLP MN**	**GMM**	**MLP VN**	**MLP MN**
Exp4	17.8	19.1	17.7	17.6	20.4	27.4
Exp5	36.7	31.4	25.9	20.3	21.9	29.4
Exp6	21.6	18.6	17.0	24.2	16.2	31.1

Table 4 GMM/MLP systems performance with telephonic speech (not completely consistent), without (NONE) and with CMN. In MLP based system UVR was used.

The results with more consistent telephonic speech (exps. 7, 8 and 9) are shown in table 5. This experiments shown a more realistic situation than the previous one.

EER(%)	NONE			CMN		
	GMM	**MLP VN**	**MLP MN**	**GMM**	**MLP VN**	**MLP MN**
Exp7	13.9	10.7	10.5	14.4	9.4	15.5
Exp8	15.2	8.8	8.6	14.3	6.7	12.3
Exp9	13.7	11.0	9.9	15.3	5.2	12.9

Table 5 GMM/MLP systems performance with telephonic speech (completely consistent), without (NONE) and with CMN. In MLP based system UVR was used.

6. Conclusions

From the results, it can be derived that GMM outperforms MLP when using microphonic speech. The differences are specially important with CMN, where GMM get the best results, and MLP the worst. An unexpected result is that, with this kind of speech, CMN seems not to work together with MLP, and we have not a clear explanation for that result, specially when this does not happen in telephonic speech with the correct normalization.

Completely different is the situation with telephonic speech. The differences are particularly relevant when the speech is obtained in a real call acquisition process (exps 7-9). For the best result (exp 9) MLP reduces GMM EER from 15.3% to 5.2% (66% less). Large EER reductions are obtained in the other two experiments, too. Besides, against what happen with GMM, the MLP performance is consistently better when the amount of training data is increased.

With regard to the tendency set out at the beginning of the work: “in the last years, the use of ANN in SR has decreased substantially”, we can conclude that the use of ANNs, as classifiers, in SR still is a interesting work line, particularly in the important field of the telephonic speech. For real applications still there are several practical problems to solve, basically related with the training stage. Solving that problems and improving the system are the principal goals of our future work.

To conclude, and focusing in the MLP based system training problems (section 4.1) and the proposed solutions, we can derive that, for our classification problem, UVR performs better than IVR; and regarding with the normalization, VN is advisable, particularly in telephonic speech.

7. References

[1] Matsui, T. and Furui, S., “Comparison of Text Independent Speaker Recognition Methods Using VQ Distortion and Discrete/Continous HMMs”, Proc. IEEE ICASSP, Vol.2, pp. 157-160, San Francisco (USA), 1992

[2] Zheng, Y. C. and Yuan B. Z., “Text Dependent Speaker Identification Using Circular Hidden Markov Models”, Prc. ICASSP, S13.3, pp. 580-582, 1988

[3] Bennani, Y. and Gallinari, P., “Connectionist Approaches for Automatic Speaker Recognition”, ESCA Workshop on Automatic Speaker Recognition Identification Verification, pp.95-102, Martigny (Switzerland), 1994.

[4] Silva, H., Vivaracho, C.E., Alonso, L. and Cardeñoso, V., “Speaker Verification: Acomparison Between ANNs and HMMs approach”, Proc Workshop on ANNs: Current Trends and Applicatio,ns. 4th World Congress on Expert Systems, Ciudad de Mexico (Mexico), 1998.

[5] Vivaracho, C.E., Alonso, L. and Moro, Q.I., “Remote Acces Control by Means of Speech”, Proc. 33rd IEEE International Carnahan Conference on Security Technology, pp.187-190, Madrid (Spain), 1999.

[6] Ortega-García, J., Cruz-Llamas, S. and Gozález-Rodríguez, J., “Facing Severe Channel Variability in Forensic Speaker Verification Conditions”, Prc. Eurospeech99, Vol. 2, pp. 783-786, Budapest (Hungary), 1999.

[7] Ortega-García, J. et al., “AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification”, IEEE ICASSP, Vol. II, pp. 773-776, 1998.

[8] Furui, S., “An Overview of Speaker Recognition Technology”, ESCA Workshop on Automatic Speaker Recognition Identification Verification, pp. 1-9, Martigny (Switzerland), 1994.

[9] Farrel, K.R., Mammone, R.J. and Assaleh, K.T., “Speaker Recognition Using Neural Networks and Conventional Classifiers”, *IEEE Transations on Speech and Audio Processing,* Vol. 2, NO. 1, part II, 1994.

[10] Kohonen, T., “Self-Organizing Maps”, *Springer,* 1997.

[11] Artieres, T., Bennani, Y. Gallinari, P. and Montacie, C., “Connectionist and Conventional Models for Free Text Talker Identification”, Neuro-Nimes, France, 1991.