Application of the Mutual Information Minimization to speaker recognition / identification improvement

Jordi Solé-Casals[1], Marcos Faundez-Zanuy[2]

[1] Signal Processing Group, University of Vic (Catalonia, Spain)
jordi.sole@uvic.es
http://www.uvic.es/eps/recerca/ca/processament/inici.html

[2] Escola Universitària Politècnica de Mataró, UPC (Catalonia, Spain)
faundez@eupmt.es

Abstract. In this paper we propose the inversion of nonlinear distortions in order to improve the recognition rates of a speaker recognizer system. We study the effect of saturations on the test signals, trying to take into account real situations where the training material has been recorded in a controlled situation but the testing signals present some mismatch with the input signal level (saturations). The experimental results for speaker recognition shows that a combination of several strategies can improve the recognition rates with saturated test sentences from 80% to 89.39%, while the results with clean speech (without saturation) is 87.76% for one microphone, and for speaker identification can reduce the minimum detection cost function with saturated test sentences from 6.42% to 4.15%, while the results with clean speech (without saturation) is 5.74% for one microphone and 7.02% for the other one.

1. Introduction

This paper proposes a non-linear channel distortion estimation and compensation in order to improve the recognition rates of a speaker recognizer. Mainly it is studied the effect of a saturation on the test signals and the compensation of this non-linear perturbation. Although common sense says that nothing can be inferred from "redundant" information data, this asserts does not state the whole possible situations or at least those cases where this kind of information can help to overcome other problems.

A well-known problem in the context of pattern recognition [1] is that a pattern recognizer trained with an insufficient number of training samples generalizes poorly when trying to classify input data. Additionally, the higher the number of model's parameters, the higher the number of training data should be. It is generally accepted [2] that using at least ten times as many training samples per class as the number of features ($n/d > 10$) is a good practice to follow in classifier design.

In some situations the use of almost redundant information can help to improve the results. An analogous naïve example easy-to-understand is the polynomial fitting to a given set of points. Figure 1 shows the interpolation of several polynomials to a set of three points. Obviously for a first, second and third degree polynomial fitting the achieved result by means of mean square error minimization can be considered satisfactory. However, for a 17^{th} polynomial degree, the problem is ill-conditioned because the number of parameters to fit is much higher than the number of available training points. Thus, although the fitted polynomial passes though the three training points, strange phenomena take place between points. This

result can be considered unsatisfactory taking into account that the range of the "y" axis spreads in a wider range. An important fact to be taken into account is that we cannot try to set up a big model that comprises a lot of parameters if the available number of training data is not enough, because recognition rates will drop instead of improve.

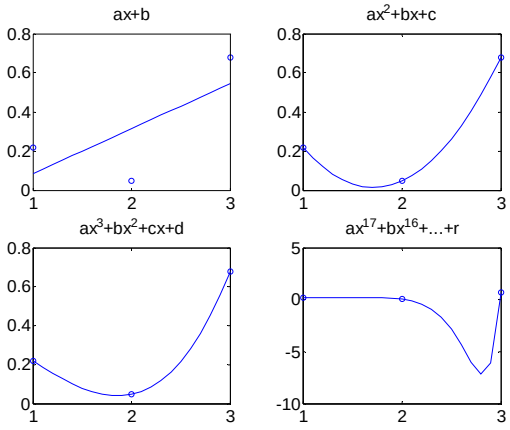

Fig. 1. Example of polynomial fitting to a set of three points.

Let us check what happens if the number of training data is artificially extended using randomly generated points, but related to the real data points. For this purpose we work out the standard deviation of the training data set:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x(i)-\bar{x}\right)^2} \tag{1}$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y(i)-\bar{y}\right)^2} \tag{2}$$

Where the experimental data set consists of N points in $\mathbb{R}^2$:

$$\left(x(i), y(i)\right) \quad i = 1, \cdots, N \tag{3}$$

And $\bar{x}, \bar{y}$ are the mean values of the x and y respectively.

The artificially generated data set $\left(x_{rand}(i), y_{rand}(i)\right) \quad i = 1, \cdots, N \times N_2$ is obtained by means of random number generation `rand`(1), which randomly generates a number on the range $[0,1]$ with a uniform distribution, using the following algorithm:

```
for i=1:N,
    for j=1:N2,
        x_rand((i-1)*N2+j)=x(i)+k*σx*(rand(1)-0.5);
        y_rand((i-1)*N2+j)=y(i)+k*σy*(rand(1)-0.5);
    end
end
```

Thus, we generate N_2 artificial points for each original one, adding a random perturbation proportional to the standard deviation of the training set.

Figure 2 shows two situations, both of them with N_2=7 ($N\times N_2$=3×7=21). The figure on the top has been obtained with a proportionality constant k=0.2, and the bottom one with k=1. It is easy to observe that in the first one the generated points are close to the original ones, while in the second case they are better distributed along the original range of signal values.

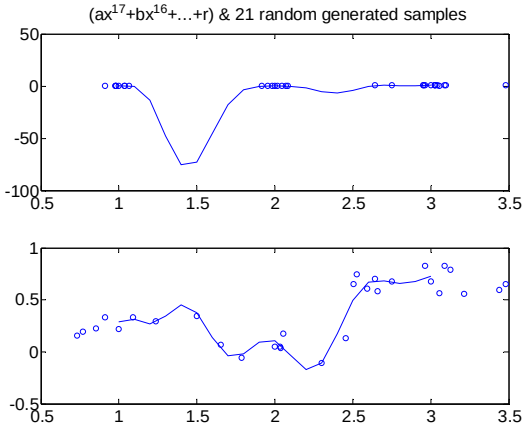

Fig. 2. Example of polynomial fitting to a set of three points plus some random generated data.

First case shown on figure 2 on the top reveals the same problem that appeared when we tried to fit the polynomial with a small experimental data set. Thus, this first example is in agreement with the initial statement "we cannot take advantage of redundant information". On the other hand, the almost-invented points on the second example produce a tight response to the original range of values.

Unfortunately, pattern recognition problems lie on higher dimensional spaces, where is not possible to plot the experimental data neither the models, so it is more complicate to understand what is really happening. However, there are experimental evidences of improvements when using redundant or almost-redundant information.

Some situations where the use of pseudo-random generated data can help to improve recognition rates are the following:

a) Pseudo-random training samples generation in order to modify the obtained statistics of the experimental data. For instance, for discriminative training, the number of inhibitory inputs is higher than the number of excitatory ones. Thus, in order to balance both amounts, one set of samples is artificially extended [3].

b) Direct modification of obtained statistics from the real experimental data. For instance, in Gaussian Mixture Models (GMM) a variance limiting constraint is used [4].

c) Replication of the known information (redundant information addition). One example is the bandwidth extension used in Digital Radio Mondiale [5].

d) Systematic generation of new training samples, theoretically "cleaner" than the original ones, and the combination of both sets of data.

In this paper we propose the inversion of nonlinear distortions in order to improve the recognition rates of a speaker recognizer system. Our proposed scheme belongs to the last category. This strategy can manage those applications where the training material has been recorded in a controlled situation but the testing signals present some mismatch with the input signal level (saturations).

By means of non-linear channel distortion estimation and compensation, we obtain a new set of feature vectors that theoretically are cleaner than the original ones. The combination of two different recognizers, one working over the original signal and another one with the compensated signal, produces an improvement on recognition rates. Figure 5 shows the proposed scheme. This approach can be interpreted as an increase on the training dataset size, or a data fusion scheme at the score's level [6]. In pattern recognition applications it is well known that a number of differently trained classifiers (that can be considered as "experts"), which share a common input, can produce a better result if their outputs are combined to produce an overall output. This technique is known as committee machine [7], ensemble averaging [8], data fusion, etc. The motivation for its use is twofold [7]:

- If the combination of experts were replaced by a single classifier, the number of equivalent adjustable parameters would be large, and this implies more training time and local minima problems [9].
- The risks of overfitting the data increases when the number of adjustable parameters is large compared to the size of the training data set.

In addition, this strategy improves the vulnerability of biometric systems [10], which is one of the main drawbacks of these systems [11].

This paper is organized as follows. Section 2 describes the Wiener model, its parameterization, and obtains the cost function based on statistical independence. Section 3 summarizes the speaker recognition/verification application. Finally, section 4 deals the experiments using the blind inversion in conjunction with the speaker recognition/verification application.

2. Non-parametric approach to blind deconvolution of nonlinear channels

When linear models fail, nonlinear models appear to be powerful tools for modeling practical situations. Many researches have been done in the identification and/or the inversion of nonlinear systems. These works assume that both the input and the output of the distortion are available [12]; they are based on higher-order input/output cross-correlation [13], bispectrum estimation [14, 15] or on the application of the Bussgang and Prices theorems [16, 17] for nonlinear systems with Gaussian inputs.

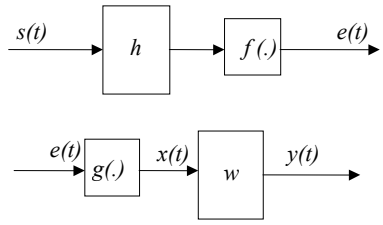

Fig. 3. The unknown nonlinear convolution system (top) and the proposed inversion structure (bottom)

However, in a real world situations, one often does not have access to the distortion input. In this case, blind identification of the nonlinearity becomes the only way to solve the problem.

This paper is concerned by a particular class of nonlinear systems, composed by a linear filter followed by a memoryless nonlinear distortion (figure 3, top). This class of nonlinear systems, also known as a Wiener system, is a nice and mathematically attracting model, but also a realistic model used in various areas [18]. We use a fully blind inversion method inspired

on recent advances in source separation of nonlinear mixtures. Although deconvolution can be viewed as a single input/single output (SISO) source separation problem in convolutive mixtures (which are consequently not cited in this paper), the current approach is actually very different. It is mainly based on equivalence between instantaneous postnonlinear mixtures and Wiener systems, provided a well-suited parameterization.

2.1 Model and assumptions

We suppose that the input of the system $S=\{s(t)\}$ is an unknown non-Gaussian independent and identically distributed (i.i.d.) process, and that subsystems h, f are a linear filter and a memoryless nonlinear function, respectively, both unknown and invertible. We would like to estimate $s(t)$ by only observing the system output. This implies the blind estimation of the inverse structure (figure 3, bottom), composed of similar subsystems: a memoryless nonlinear function g followed by a linear filter w. Such a system is known as a Hammerstein system. Let $\mathbf{s}$ and $\mathbf{e}$ be the vectors of infinite dimension, whose t-th entries are $s(t)$ or $e(t)$, respectively. The unknown input-output transfer can be written as:

$$\boldsymbol{e} = f(\mathbf{H}\boldsymbol{s}) \tag{4}$$

where:

$$\mathbf{H} = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & h(t+1) & h(t) & h(t-1) & \cdots \\ \cdots & h(t+2) & h(t+1) & h(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{5}$$

is an infinite dimension Toeplitz matrix which represents the action of the filter h to the signal s(t). The matrix $\mathbf{H}$ is non-singular provided that the filter h is invertible, i.e. satisfies $h^{-1}(t) * h(t) = h(t) * h^{-1}(t) = \delta(t)$, where δ(t) is the Dirac impulse. The infinite dimension of vectors and matrix is due to the lack of assumption on the filter order. If the filter h is a finite impulse response (FIR) filter of order N_h, the matrix dimension can be reduced to the size N_h. In practice, because infinite-dimension equations are not tractable, we have to choose a pertinent (finite) value for N_h. Equation (1) corresponds to a post-nonlinear (pnl) model [19]. This model has been recently studied in nonlinear source separation, but only for a finite dimensional case. In fact, with the above parameterization, the i.i.d. nature of s(t) implies the spatial independence of the components of the infinite vector s. Similarly, the output of the inversion structure can be written $\boldsymbol{y} = \mathbf{W}\boldsymbol{x}$ with $x(t) = g(e(t))$. Following [19, 20] the inverse system (g, w) can be estimated by minimizing the output mutual information, i.e. spatial independence of $\mathbf{y}$ which is equivalent to the i.i.d. nature of y(t), as can be seen in figure 4.

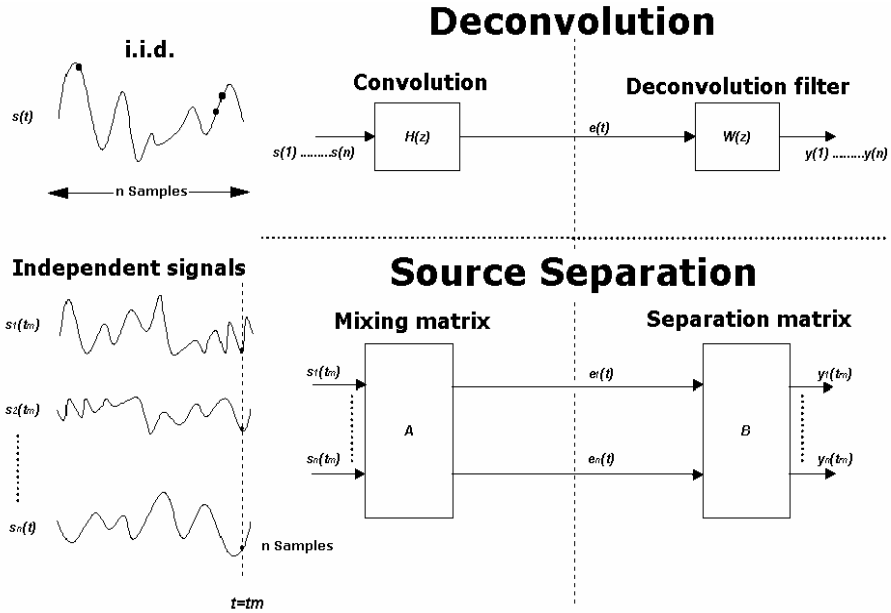

Fig. 4. Relationship between blind (linear) decnvolution and blind source separation. The spatial independence criteria used in source separation context is transformed in temporal independence in the deconvolution context.

2.2 Cost function

The mutual information of a random vector of dimension *n*, defined by

$$I(Z) = \sum_{i=1}^{n} H(z_i) - H(z_i, z_2, ..., z_n) \tag{6}$$

can be extended to a vector of infinite dimension, using the notion of *entropy rates* of stationary stochastic processes [21]:

$$I(Z) = \lim_{T\to\infty} \frac{1}{2T+1} \left\{ \sum_{t=-T}^{T} H(z(t)) - H(z(-T),...,z(T)) \right\} = H(z(\tau)) - H(Z) \tag{7}$$

where τ is arbitrary due to the stationarity assumption. We can notice that $I(Z)$ is always positive and vanishes iff $z(t)$ is i.i.d. Since S is stationary, and h and w are time-invariant filters, then Y is stationary too, and $I(Y)$ is defined by:

$$I(Y) = H(y(\tau)) - H(Y) \tag{8}$$

Using the Lemma 1 of [20], the last right term of equation (5) becomes:

$$H(Y) = H(X) + \frac{1}{2\pi} \int_0^{2\pi} \log \left| \sum_{t=-\infty}^{+\infty} w(t) e^{-jt\theta} \right| d\theta \tag{9}$$

Moreover, using $x(t) = g(e(t))$ and the stationarity of $E = \{e(t)\}$:

$$H(X) = \lim_{T\to\infty} \frac{1}{2T+1}\left\{ H(e(-T),\ldots,e(T)) + \sum_{t=-T}^{T} E[\log g'(e(t))] \right\} = H[E] + E[\log g'(e(\tau))] \quad \textbf{(10)}$$

Combining (6) and (7) in (5) leads finally to:

$$I(Y) = H(y(\tau)) - \frac{1}{2\pi}\int_0^{2\pi} \log\left| \sum_{t=-\infty}^{+\infty} w(t) e^{-jt\theta} \right| d\theta - E[\log g'(e(\tau))] - H[E] \quad \textbf{(11)}$$

3. Speaker recognition/verification

One of the main sources of degradation in speaker recognition is the mismatch between training and testing conditions. For instance, in [22] we evaluated the relevance of different training and testing languages, and in [23] we also studied other mismatch, such as the use of different microphones. In this paper, we study a different source of degradation: different input level signals in training and testing. Mainly we consider the effect of saturation. We try to emulate a real scenario where a person speaks too close to the microphone or to loud, producing a saturated signal. Taking into account that the perturbations are more damaging when they are present just during training or testing but not in both situations, we have used a clean database and artificially produced saturation in the test signals. Although it would be desirable to use a "real" saturated database, we don't have this kind of database, and the simulation give us more control about "how the algorithm is performing". Anyway, we have used a real saturated speech sentence in order to estimate the nonlinear distortion using the algorithm described in section 2 and the results have been successful. Figure 5 shows a real saturated speech frame and the corresponding estimate of the NL perturbation.

3.1 Database

For our experiments we have used a subcorpora of the Gaudi database, that follows the design of [24]. It consists on 49 speakers acquired with a simultaneous stereo recording with two different microphones (AKG C-420 and SONY ECM66B). The speech is in wav format at fs=16 kHz, 16 bit/sample and the bandwidth is 8 kHz. We have applied the potsband routine that can be downloaded from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html in order to obtain narrow-band signals. This function meets the specifications of G.151 for any sampling frequency. The speech signals are pre-emphasized by a first order filter whose transfer function is $H(z)=1-0.95z^{-1}$. A 30 ms Hamming window is used, and the overlapping between adjacent frames is 2/3. One minute of read text is used for training, and 5 sentences for testing (each sentence is about two seconds long).

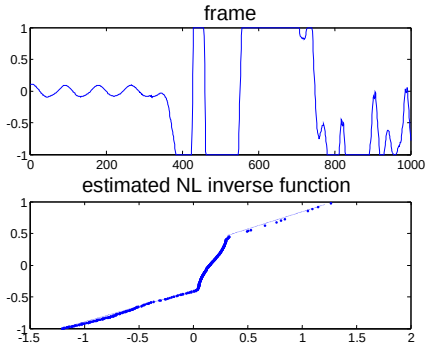

Fig. 5. Saturated frame and the estimated channel function

3.2 Speaker recognition / verification algorithm

We have chosen a second-order based measure for the recognition of a speaker. In the training phase, we compute for each speaker empirical covariance matrices (CM) based on feature vectors extracted from overlapped short time segments of the speech signals, i.e., $C_j = \hat{E}\left[x_n x_n^T\right]$, where $\hat{E}$ denotes estimate of the mean and x_n represents the features vector for frame n. As features representing short time spectra we use mel-frequency cepstral coefficients. In the speaker-recognition system, the trained covariance matrices (CM) for each speaker are compared to an estimate of the covariance matrix obtained from a test sequence from a speaker. An arithmetic-harmonic sphericity measure is used in order to compare the matrices [25]: $d = \log\left(\mathrm{tr}(C_{test} C_j^{-1})\,\mathrm{tr}(C_j C_{test}^{-1})\right) - 2\log(l)$, where $\mathrm{tr}(\cdot)$ denotes the trace operator, l is the dimension of the feature vector, C_{test} and C_j is the covariance estimate from the test speaker and speaker model j, respectively. In the speaker-verification system, the algorithm is basically the previous one, were have applied the following equation in order to convert the distance measure d into a probability measure p: $p = e^{-0.5d}$, and the system has been evaluated using the DET curves [26], with the following detection cost function (DCF): $DCF = C_{miss} \times P_{miss} \times P_{true} + C_{fa} \times P_{fa} \times P_{false}$ where C_{miss} is the cost of a miss, C_{fa} is the cost of a false alarm, P_{true} is the a priori probability of the target, and $P_{false} = 1 - P_{true}$. We have used $C_{miss} = C_{fa} = 1$. Figure 6 shows an example of DET plot.

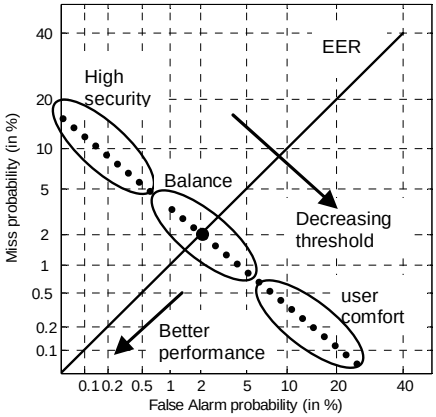

Fig. 6. Example of a DET plot for a speaker verification system (dotted line). The Equal Error Rate (EER) line shows the situation where False Alarm equals Miss Probability (balanced performance). Of course one of both errors rates can be more important (high security application versus those where we do not want to annoy the user with a high rejection/ miss rate). If the system curve is moved towards the origin, smaller error rates are achieved (better performance). If the decision threshold is reduced, we get higher False Acceptance/Alarm rates.

4. Experiments and conclusions

Using the database described in section 3, we have artificially generated a test signal database, using the following procedure:

- All the test signals are normalized to achieve unitary maximum amplitude.
- A saturated database has been artificially created using the following equation:
- $x' = tanh(kx)$, where k is a positive constant.

The training set remains the same, so no saturation is added. In order to show the improvement due to the compensation method, figure 6 shows one frame that has been artificially saturated with a dramatic value (k=10), the original, and the recovered frame applying the blind inversion of the distortion.

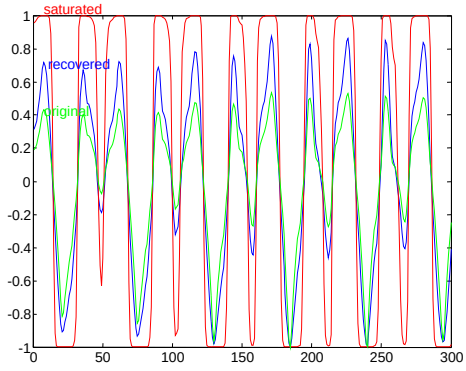

Fig. 7. Example of original, saturated, and recovered frame using the proposed procedure.

Using the original (clean) and artificially generated database (saturated) we have evaluated the identification rates and the minimum DCF. For the saturated test sentences scenario, we have estimated one different channel model for each test sentence, applying the method described in section 2. This is a way to manage real situations where the possible amount of saturation is not known in advance and must be estimated for each particular test sentence. In order to improve the results an opinion fusion is done, using the scheme shown in figure 7. Thus, we present the results in three different combination scenarios for speaker recognition:

- Just one opinion (1 or 2 or 3 or 4)
- To use the fusion of two opinions (1&2 or 2&3).
- The combination of the four available opinions.

Table 1, for speaker recognition experiments, and Table 2, for speaker verification experiments, show the results for k=2 in all this possible scenarios using two different combinations [6] rules (arithmetic and geometric mean, [27]), with a previous distance normalization [28].

Table 1. Results for several classifiers, shown in figure 7.

Combination		Recognition rate
1 (AKG+NL compensation)		83.67 %
2 (AKG)		82.04 %
3 (SONY+NL compensation)		80.82 %
4 (SONY)		80 %
1&2	Arithmetic	84.9 %
	Geometric	84.9 %
1&3	Arithmetic	89.39%
	Geometric	87.35%
2&4	Arithmetic	88.16%
	Geometric	86.53 %
1&2&3&4	Arithmetic	88.16 %
	Geometric	87.76 %

Table 2. Minimum Detect Cost Function for several classifiers, shown in figure 7.

Combination		Minimum DCF
1 (AKG+NL compensation)		6.42 %
2 (AKG)		5.74 %
3 (SONY+NL compensation)		6.59 %
4 (SONY)		7.02 %
1&2	Arithmetic	5.95 %
	Geometric	5.95 %
1&3	Arithmetic	4.15 %
	Geometric	4.89 %
3&4	Arithmetic	6.99 %
	Geometric	6.21 %
2&4	Arithmetic	4.61 %
	Geometric	5.53 %
1&2&3&4	Arithmetic	4.43 %
	Geometric	5 %

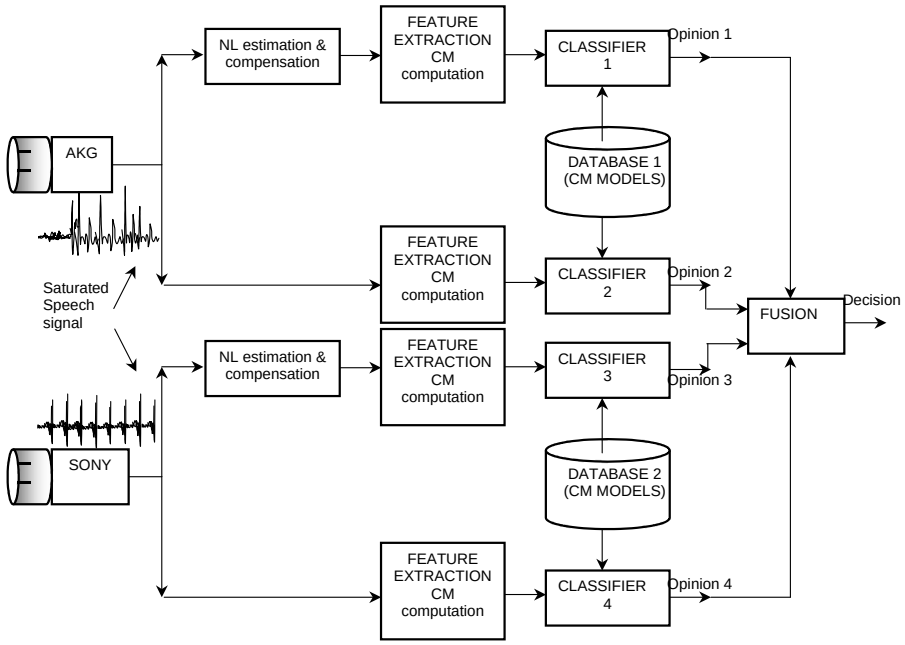

Fig. 8. General Scheme of the recognition system

Main conclusions are:

- The use of the NL compensation improves the obtained results with the same conditions than without this compensation block.
- The combination between different classifiers improves the results. These results can be even more improved using a weighted sum instead a mean. Anyway, we have preferred a fixed combination rule than a trained rule.
- We think that using a more suitable parameterization, the improvements would be higher.

Acknowledgments

This work has been supported by COST action 277, University of Vic under the grant R0912, FEDER & CICYT TIC-2003-08382-C05-02

References

[1] Jain A. K., Duin R. P. W., Mao J., "Statistical pattern recognition: a review". IEEE trans. On Pattern Analysis and Machine Intelligence. Vol.22, No 1, January 2000.

[2] Jain A. K., Chandrasekaran B., "Dimensionality and sample size considerations in pattern recognition practice", Handbook of statistics. P. R. Krishnaiah and L. N. Kanal, eds. Vol. 2, pp. 835-855, Amsterdam: North-Holland 1982.

[3] Faundez-Zanuy M., “Data fusion in biometrics” Accepted for publication, IEEE Aerospace and Electronic Systems Magazine. In press, 2005.
[4] Bishop C.M. “Neural networks for pattern recognition” Ed. Clarendon press. 1995
[5] Reynolds D. A., Rose R. C., “Robust text-independent speaker identification using gaussian mixture speaker models”. IEEE Trans. On speech and audio processing, Vol.3, No 1, pp. 72-83, January 1995
[6] http://www.drm.org
[7] Haykin S., Chapter 7, Committee Machines. “Neural nets. A comprehensive foundation”, 2on edition. Ed. Prentice Hall 1999
[8] Perrone M. P., and Cooper L. N. “When networks disagree: ensemble methods for hybrid neural networks” in “neural networks for speech and image processing, R. J. Mammone ed., Chapman-Hall 1993
[9] Jain A. K., and Mao J., “Artificial neural networks: a tutorial”. IEEE Computer pp.31-44, March 1996
[10] Faundez-Zanuy M., “On the vulnerability of biometric security systems”. IEEE Aerospace and Electronic Systems Magazine. Vol.19 nº 6, pp.3-8, June de 2004.
[11] Faundez-Zanuy M., “Biometric recognition: why not massively adopted yet?” Accepted for publication, IEEE Aerospace and Electronic Systems Magazine. In press, 2005.
[12] S. Prakriya, D. Hatzinakos. Blind identification of LTI-ZMNL-LTI nonlinear channel models. Biol. Cybern., 55 pp. 135-144 (1985).
[13] S.A. Bellings, S.Y. Fakhouri. Identification of a class of nonlinear systems using correlation analysis. Proc. IEEE, 66 pp. 691-697 (1978).
[14] C.L. Nikias, A.P. Petropulu. Higher-Order Spectra Analysis – A Nonlinear Signal processing Framework. Englewood Cliffs, NJ: Prentice-Hall (1993).
[15] C.L. Nikias, M.R.Raghuveer. Bispectrum estimation: A digital signal processing framework. Proc. IEEE, 75 pp. 869-890 (1987)
[16] E.D. Boer. Cross-correlation function of a bandpass nonlinear network. Proc. IEEE, 64 pp. 1443-1444 (1976)
[17] G. Jacoviti, A. Neri, R. Cusani. Methods for estimating the autocorrelation function of complex stationary process. IEEE Trans. ASSP, 35, pp. 1126-1138 (1987)
[18] J. Solé, C. Jutten, A. Taleb “Parametric approach to blind deconvolution of nonlinear channels”. Ed. Elsevier, Neurocomputing 48 pp.339-355, 2002
[19] A. Taleb, C. Jutten. Source separation in postnonlinear mixtures. IEEE Trans. on S.P., Vol. 47, n°10, pp.2807-20 (1999).
[20] A. Taleb, J. Solé, C. Jutten. Quasy-Nonparametric Blind Inversion of Wiener Systems. IEEE Trans. on S.P., Vol. 49, n°5, pp.917-924 (2001).
[21] T.M. Cover, J.A. Thomas. Elements of Information Theory. Wiley Series in Telecommunications (1991)
[22] A. Satué, M. Faúndez-Zanuy "On the relevance of language in speaker recognition" EUROSPEECH 1999 Budapest, Vol. 3 pp.1231-1234
[23] C. Alonso, M. Faúndez-Zanuy, “Speaker identification in mismatch training and testing conditions”.Vol. II, pp. 1181-1184, IEEE ICASSP’2000, Istanbul
[24] J. Ortega, J. Gonzalez & V. Marrero, “Ahumada: a large speech corpus in spanish for speaker characterization and identification”, Speech Communication 31, pp.255-264, 2000.
[25] F. Bimbot, L. Mathan "Text-free speaker recognition using an arithmetic-harmonic sphericity measure." pp.169-172, Eurospeech 1993.
[26] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection performance”, V. 4, pp.1895-1898, Eurospeech 1997
[27] J. Kittler, M. Hatef, R. P. W. Duin & J. Matas “On combining classifiers”. IEEE Trans. On pattern analysis and machine intelligence, Vol. 20, Nº 3, pp. 226-239, march 1998.
[28] C. Sanderson “Information fusion and person verification using speech & face information”. IDIAP Research Report 02-33, pp. 1-37. September 2002