# On the Use of Jacobian Adaptation in Real Speaker Verification Applications

*Jan Anguita and Javier Hernando*

TALP Research Center
Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
Barcelona, Spain
{jan,javier}@talp.upc.es

## Abstract

Jacobian Adaptation (JA) of the acoustic models is a fast adaptation technique that has been successfully used in both speech and speaker recognition systems. This technique adapts the acoustic models on the basis of the difference between the testing and the training noise conditions. For this reason, a noise reference of both the training and the testing phase is needed. In previous works, the noise conditions have been commonly supposed to be known, or estimated from the first part of the signals or using manually obtained *speech* and *non-speech* labels. However, in a real application the noise conditions are not generally known, it is not sure that the first part of the signal does not contain speech and manual labels are not usually available. In this work we propose to obtain the noise references by using continuous noise estimation methods, which are more appropriate for real applications. Several noise estimation methods are compared and the obtained results show that these techniques are effective for JA.

**Index Terms**: speaker verification, Jacobian adaptation, noise estimation

## 1. Introduction

Speaker recognition systems are severely degraded by mismatch between training and testing conditions. Several techniques such as speech enhancement, novel speech parameterizations and model adaptation have been proposed in the last years in order to reduce the influence of this mismatch and improve the recognition results.

Jacobian Adaptation (JA) is a model adaptation technique used to adapt a set of models from certain noise conditions to other conditions in order to alleviate the mismatch between training and testing [1-5]. This technique adapts the models on the basis of the difference between the training and the testing noise conditions. Therefore, it is necessary to estimate a noise reference of both the training and the testing phase. Obviously, the more accurate is the noise estimation the better should be the adaptation of the models.

In our previous works it has been shown that JA can improve the results of speaker recognition systems in noisy conditions [4][5]. In these works, the training set of the database was manually labeled and the *speech* and *non-speech* labels were used to estimate the speaker models and the noise references of the training phase. In the testing phase the noise reference was obtained by averaging a constant number of frames at the beginning of the signals, assuming that these frames do not contain speech. The objective of the work presented in this paper is to study how to use JA in more realistic applications where manual labels are not available and it is not always true that a segment of fixed length at the beginning the signals does not contain speech.

In order to deal with the lack of labels problem we consider two training conditions: train the speakers with the whole signals and use a speech activity detector (SAD) during the training phase. There are several proposals in the literature to deal with the noise estimation problem in JA. In [1], the noise conditions are supposed to be known, which is not generally true in a real situation. In [2] the noise estimation is obtained by averaging the *N* frames of the signal with the lowest energy but this procedure biases the estimation toward lower values. In [3][4] the noise is estimated by averaging a constant number of frames at the beginning of the signals.

None of noise estimation methods commented above are appropriate for a real application. For this reason, in this work we propose to use a continuous noise estimation method to obtain the noise references. These methods have the advantage that no previous information about the environment is needed. In addition, the entire signals can be used to estimate the noise, which should provide a better estimation in case of non-stationary noises. A SAD could also be used to estimate the noise from the silence parts, but previous works have shown that continuous noise estimation methods provide more accurate results [6] and sometimes there is not enough silence to estimate the noise. Three noise estimation methods are compared in this work: Recursive Averaging (RA) [7], Minimum Statistics (MS) [6] and Minima Controlled Recursive Averaging (MCRA) [8].

In section 2 we review the JA technique used in this work. In section 3 the noise estimation methods are described. In section 4 the experimental results are presented. Finally, section 5 contains the conclusions of this work.

## 2. Model-dependent noise reference Jacobian adaptation

Jacobian Adaptation is an acoustic model compensation technique used to adapt the mean vectors of Hidden Markov Models (HMM) from certain noisy conditions to other conditions [1]. The JA transformation is as follows:

$$\hat{\mathbf{C}}_{s+n} = \mathbf{C}_{s+n} + \frac{\partial \mathbf{C}_{s+n}}{\partial \mathbf{C}_n}\left(\hat{\mathbf{C}}_n - \mathbf{C}_n\right) \quad (1)$$

where $\partial \mathbf{C}_{s+n}/\partial \mathbf{C}_n$ is the Jacobian matrix, $\hat{\mathbf{C}}_{s+n}$ and $\mathbf{C}_{s+n}$ are the new adapted and the original noisy speech feature mean vector respectively, and $\mathbf{C}_n$ and $\hat{\mathbf{C}}_n$ are the reference and the target noise feature vector, i.e., the noise present in the training signals and the actual noise present in the recognition phase respectively. In [4] it is shown that JA is more effective for speaker verification when Frequency Filtering (FF) [9] parameters are used. In this case, the Jacobian matrix can be written as [3]:

$$\frac{\partial \mathbf{C}_{s+n}}{\partial \mathbf{C}_n} = \mathbf{H} diag\left(\frac{\alpha \mathbf{N}}{\mathbf{S} + \alpha \mathbf{N}}\right)\mathbf{H}^{-1} \quad (2)$$

where **N** is the filter-bank energies (FBE) vector of the training noise reference, **S** is the FBE vector of the noisy speech model, **H** is the FF matrix transformation and $\mathbf{H}^{-1}$ is its inverse [9]. The quotient is computed element by element and *diag()* is the diagonal matrix formed with elements of the vector inside. The constant parameter $\alpha$ is included to alleviate the problem of JA for large mismatch between training and testing conditions [2].

In general in JA only one noise reference was used to calculate all the Jacobian matrices and to adapt all the models [1-3]. In [4] we proposed a modification of JA, which was called Model-dependent Noise Reference Jacobian Adaptation (MNRJA), where a different noise reference is used to adapt each model. This technique was shown to reduce the speaker verification error rates because in speaker recognition systems the utterances of a speaker are not used to train the models of other speakers. In the training phase, a noise reference is estimated for each speaker from the speech signals used to train his/her model. Then, the Jacobian matrices (one for each mean vector of each mixture of each model) are calculated using equation (2), where the vector **N** used in is the specific noise reference FBE of each model. In the testing phase the models are adapted using equation (1), where the vector $\mathbf{C}_n$ is the specific noise reference cepstrum vector of the model to adapt.

## 3. Noise estimation

The JA algorithm described in the previous section relies on a noise estimation method both in the training and the testing phase. The used method has to be accurate and practical for a real application. Previous works have obtained good results with JA but the used noise estimation methods are not appropriate for a real application. In [1] the noise during the testing phase is supposed to be known, which is not true in a real situation. In [2] the noise is estimated by averaging the 10 frames of the signal with the lowest energy, which is a simplification of the Minimum Statistics technique considered in this work. In addition, since the minimum of a non trivial random variable is always lower than the average, this estimation is biased toward lower values. In [3][4] the noise is obtained from the average of the first segment of the testing signal. This method has the drawback that it is not easy to determine the optimal length of the segment used to estimate the noise. In fact, it is not sure that a fixed interval of the testing signal does not contain speech. In addition, if the noise is non-stationary, the reference is not representative of the entire signal. For these reasons we propose to use a continuous noise estimation method. In this way, the entire testing signal can be used to estimate the noise and no information about the intervals where speech is absent is needed.

The objective of the noise estimation module in JA is to calculate the noise reference vectors $\mathbf{C}_n$ and **N** during the training phase and the vector $\hat{\mathbf{C}}_n$ during the testing phase. During the training phase the noise spectrum of each signal is estimated using a continuous noise estimation method. After this, the parameterized FF vectors are calculated for each frame. The training noise reference $\mathbf{C}_n$ is obtained by averaging over all the frames. The averaging is done first over each signal and then over all the signals. Finally, the noise reference in the FBE domain **N** is calculated from $\mathbf{C}_n$. The procedure in the testing phase is the same but using the testing signal. In the next sections, the noise estimation methods used in this work are described.

### 3.1. Recursive averaging (RA)

This method calculates the noise estimation as follows [7]:

$$N(l,k) = \begin{cases} \alpha_R N(l-1,k) + (1-\alpha_R) Y(l,k) & \text{if } \dfrac{Y(l,k)}{N(l-1,k)} < \beta \\ N(l-1,k) & \text{otherwise} \end{cases} \quad (3)$$

where *Y(l,k)* is the noisy signal spectral magnitude at time *l* and frequency bin *k*, *N(l,k)* is the noise magnitude estimation, $\alpha_R$ is the smoothing parameter of the recursive averaging ($0 < \alpha_R < 1$) and $\beta$ is a threshold that controls the recursion. This technique calculates the noise estimation as a weighted sum of past spectral values *Y(l,k)* in each frequency bin *k* when *Y(l,k)* does not contain speech. When *Y(l,k)* is higher than $\beta N(l-1,k)$ it is considered that speech is present and the recursion is stopped.

### 3.2. Minimum statistics (MS)

This method is based on the assumption that speech and noise are independent and that the power of a noisy speech signal frequently decays to the power level of the noise [6]. The noise estimation is obtained from the smoothed periodogram calculated as follows:

$$P(l,k) = \alpha_M(l,k) P(l-1,k) + (1-\alpha_M(l,k)) Y^2(l,k) \quad (4)$$

where *P(l,k)* is the smoothed periodogram and $\alpha_M(l,k)$ is the smoothing parameter of the recursive averaging. The noise power estimation $N^2(l,k)$ is obtained by taking the minimum value of the $D_M$ last frames of *P(l,k)*:

$$N^2(l,k) = \min\{P(l',k) \mid l - D_M + 1 \le l' \le l\} \quad (5)$$

The smoothing parameter $\alpha_M(l,k)$ is calculated to minimize the mean square error between the smoothed periodogram *P(l,k)* and the estimated noise $N^2(l,k)$. In [6] it is shown that this is accomplished when:

$$\alpha_M(l,k) = \frac{\alpha_{\max}}{1 + \left(\dfrac{P(l-1,k)}{N^2(l-1,k)} - 1\right)^2} \quad (6)$$

where $\alpha_{max}$ is the maximum value of $\alpha_M(l,k)$. Since the noise is estimated from the minimum of the periodogram, the estimator is biased. For this reason, the final noise estimation is multiplied by a bias correction factor $B_{min}(l,k)$. In addition, an error monitoring strategy is used in order to detect tracking errors in the $P(l,k)$ estimate. For the details on the calculation of $B_{min}(l,k)$ and the error monitoring method please refer to [6].

### 3.3. Minima controlled recursive averaging (MCRA)

This technique calculates the noise estimation as follows [8]:

$$N^2(l,k) = \alpha_C(l,k)N^2(l-1,k) + (1-\alpha_C(l,k))Y^2(l,k) \quad (7)$$

where $\alpha_C(l,k)$ is the smoothing parameter of the recursive averaging. This parameter is controlled by the speech presence probability $p(l,k)$ as follows:

$$\alpha_C(l,k) = \alpha_d + (1-\alpha_d)\,p(l,k) \quad (8)$$

where $\alpha_d$ is a constant ($0 < \alpha_d < 1$). In order to calculate the speech presence probability the power spectrum of the noisy signal is smoothed in time and frequency:

$$S_f(l,k) = \sum_{i=-w}^{w} b(i)Y^2(l,k-i) \quad (9)$$

$$S(l,k) = \alpha_S S(l-1,k) + (1-\alpha_S)S_f(l,k) \quad (10)$$

where $\alpha_S$ is a constant ($0 < \alpha_S < 1$) and $b(i)$ is a Hanning window of size $2w+1$. The minimum value of $S(l,k)$ in the last $D_C$ frames $S_{min}(l,k)$ is used to compute the ratio $S_r(l,k)=S(l,k)/S_{min}(l,k)$. It is decided that speech is present at frame $l$ and frequency bin $k$ if $S_r(l,k) > \delta$. The speech presence probability is calculated as:

$$p(l,k) = \alpha_p p(l,k) + (1-\alpha_p)I(l,k) \quad (11)$$

where $\alpha_p$ is a constant ($0 < \alpha_p < 1$) and $I(l,k) = 1$ if $S_r(l,k) > \delta$ and $I(l,k) = 0$ otherwise. This formulation takes into account the strong correlation of speech presence in consecutive frames.

## 4. Experiments and results

### 4.1. Training conditions

In our previous works it has been shown that JA can improve the results of speaker recognition systems in noisy conditions [4][5]. In these works the training set of the database was manually labeled in order to know which segments contained speech and which ones contained non-speech sounds. In the training phase, these labels were used to train the speaker models from the speech segments, and to estimate the training noise reference for JA and a silence model from the non-speech segments. In the testing phase the noise was estimated by averaging a constant number of frames at the beginning of the signals. Although this testing framework is useful to evaluate the effectiveness of JA for speaker verification, it is interesting to study how to apply this technique in a more realistic and restrictive application. For this reason, in this work we do not use manual labels and we evaluate two different training conditions: without labels and obtaining the labels of the training signals with a SAD [10].

- **Without labels**: the speaker models are trained using the whole signals. The noise references for JA are estimated by averaging the first part of the signals (in this case this part is not used to train the speaker model) or with a continuous noise estimation method.
- **SAD**: the *speech* and *non-speech* labels are used to train the speaker models and a silence model. The noise references for JA are estimated either by averaging the *non-speech* parts of the signals in the training phase and averaging the first part of the signals in the testing phase or with a continuous noise estimation method.

### 4.2. Experimental setup

Experimental evaluation was carried out with the BioTech database, which is a private multi-session speaker-oriented database recorded with an 8kHz sampling rate. It contains 184 speakers, 106 males and 78 females, which were recorded using both land-line and mobile telephones. In this evaluation we perform text independent experiments by using signals containing four random digits (two utterances per session). Noise was added to the original BioTech database to obtain four different noise types ('PC', 'Bar', 'Keyboard' and 'Factory') at four SNRs (30, 20, 10 and 0dB).

A client model per speaker was built from the first four sessions of the 100 speakers that have more that four sessions. A universal background model (UBM) was built from the 33 speakers that only have one session. The speakers that have between two and four sessions are used as impostors. True-identity tests were made on the 5$^{th}$ and later sessions. With existing sessions for 100 client speakers chosen, this gives 578 true-identity tests. A total of 1605 false-identity tests were made with no cross-gender tests.

The verification system was implemented with HTK modified to include the FF representation. In the parameterization stage, the speech signal was divided into frames of 30ms at a rate of 10ms. Each frame was characterized with 20 FF parameters plus the first and second time derivatives. Each speaker and the UBM were characterized by a Gaussian Mixture Model with 32 mixtures. The silence was modeled by a Hidden Markov Model of 3 states with one Gaussian per state. JA of only static components of mean vectors was implemented. The $\alpha$ parameter of JA was set to 3. The values of the RA parameters were: $\alpha_R = 0.95$ and $\beta = 2$. The values of the MS parameters were: $\alpha_{max} = 0.95$ and $D_M = 120$. The values of the MCRA parameters were: $\alpha_d = 0.95$, $\alpha_S = 0.8$, $\alpha_p = 0.2$, $\delta = 5$, $D_C = 120$, $w = 1$.

### 4.3. Experimental results

A complete set of experiments has been carried out without adaptation and with JA. Training for each technique was done with PC noise environment at 30dB of SNR. All test signals (four noise types at four different SNRs) were used to test each technique.

Tables 1 and 2 show the average EER (averaging over all noise types and SNRs) obtained without adaptation (FF) and with JA estimating the noise with different techniques:

averaging the first part of the signals (125, 250, 500 and 1000ms considering that the signals duration is around 4s) and with a continuous noise estimation method. In these tables, we can observe that the error rate is lower when using a SAD during the training phase if JA is not used. The difference in terms of EER between the two training conditions (SAD or without labels) is reduced when using JA. Table 1 shows that JA can improve the performance of the system if the length of the segment used to estimate the noise references correctly chosen. Unfortunately, the optimal length depends on the training condition. In addition, with other evaluation databases the optimal length can change. On the other hand, Table 2 shows that, when the noise references are estimated with a continuous noise estimation method, JA improves the results in both training conditions. Moreover, the EERs obtained with the RA and MCRA methods are lower than those shown in Table 1.

| | | JA | | | |
|---|---|---|---|---|---|
| | **FF** | **125ms** | **250ms** | **500ms** | **1s** |
| **SAD** | 16.49 | 20.23 | 19.84 | 18.96 | 15.32 |
| **Without lab.** | 18.67 | 16.20 | 15.87 | 19.12 | 20.10 |

Table1: *Average EER when the noise estimated by averaging the first segment of the signals.*

| | | JA | | |
|---|---|---|---|---|
| | **FF** | **RA** | **MS** | **MCRA** |
| **SAD** | 16.49 | 13.60 | 16.13 | 13.51 |
| **Without lab.** | 18.67 | 14.20 | 16.00 | 15.22 |

Table 2: *Average EER when the noise estimated with continuous noise estimation methods.*

Tables 1 and 2 show that the best results are obtained by using a SAD to train the speaker models, with JA and estimating the noise references with RA or MCRA. Tables 3 and 4 show the average EERs for each noise type and each SNR when the noise is estimated with the RA method, which has a lower computational cost than MCRA. These tables also show the relative improvement obtained with adaptation. Table 3 shows that the lowest EERs are obtained with the PC and Factory noises, which are the most stationary ones. The higher EER reduction when using JA is obtained with PC noise. Table 4 shows that the error rates increase when the SNR decreases. It also shows that the improvement is mainly obtained at low SNRs (0 and 10dB). When the SNR is high (20 or 30dB), the results are almost the same with and without adaptation.

| | **FF** | **JA-RA** | **Rel. Imp.** |
|---|---|---|---|
| **PC** | 15.15 | 10.77 | 28.92% |
| **Bar** | 19.52 | 15.95 | 18.30% |
| **Factory** | 12.16 | 10.55 | 13.24% |
| **Keyboard** | 19.14 | 17.13 | 10.49% |

Table 3: *Relative EER reduction obtained with JA-RA averaging over the testing SNRs.*

| | **FF** | **JA-RA** | **Rel. Imp.** |
|---|---|---|---|
| **SNR 30dB** | 10.57 | 10.47 | 1.00% |
| **SNR 20dB** | 11.04 | 11.07 | -0.27% |
| **SNR 10dB** | 17.06 | 13.57 | 20.42% |
| **SNR 0dB** | 27.30 | 19.29 | 29.35% |

Table 4: *Relative EER reduction obtained with JA-RA averaging over the testing noise types.*

## 5. Conclusions

JA adapts the acoustic models on the basis of the difference between the testing and the training noise conditions. For this reason a noise reference of both the training and the testing phase is needed. In this work we have studied the use of JA in a speaker verification application where the training signals are not manually labeled and no information about the testing signals is available. In order to estimate the noise references we have used continuous noise estimation methods. The obtained results show that these methods, besides being more appropriate for real application than previous proposals, obtain lower speaker verification error rates.

## 6. References


[1] Sagayama S., Yamaguchi Y., Takahashi S. and Takahashi J., "Jacobian Approach to Fast Acoustic Model Adaptation", Proceedings of the ICASSP, vol. 2, pp. 835-838, 1997.

[2] Cerisara C., Rigazio L., Boman R. and Junqua J.-C., "α-Jacobian environmental adaptation", Speech Communication, vol. 42, pp. 25-41, 2004.

[3] Abad A., Nadeu C., Hernando J. and Padrell J., "Jacobian Adaptation based on the Frequency Filtered Spectral Energies", Proceedings of the Eurospeech, 2003.

[4] Anguita J., Hernando J. and Abad A., "Improved Jacobian Adaptation for Robust Speaker Verification", IEICE Transactions on Information and Systems, Vol. E88-D, No. 7, pp. 1767-1770, July 2005.

[5] Anguita J., Hernando J. and Abad A., "Jacobian Adaptation with Improved Noise Reference for Speaker Verification", Proceedings of the ICSLP, 2004.

[6] Martin R., "Noise Power Spectral density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5, pp. 504-512, July 2001.

[7] Hirsch H. G. and Ehrlicher C., "Noise Estimation Techniques for Robust Speech Recognition", Proceedings of the ICASSP, pp. 153-156, 1995.

[8] Cohen I. and Berdugo B., "Speech Enhancement for Non-stationary Noise Environments", Signal Processing, Vol. 81, pp. 2403-2418, 2001.

[9] Nadeu C., Macho D. and Hernando J., "Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition", Speech Communication, vol. 34, pp. 93-114, 2001.

[10] Padrell J., Macho D. and Nadeu C., "Robust Speech Activity Detection Using LDA Applied to FF Parameters", Proceedings of the ICASSP, pp. 557-560, 2005.