BLIND EQUALIZATION FOR ROBUST TELEPHONE BASED SPEECH RECOGNITION

Laurent MAUUARY
e-mail: mauuary@lannion.cnet.fr
France Télécom, Centre National d'études des télécommunications, CNET/LAA/TSS/RCP,
Technopole Anticipa, 2, avenue Pierre Marzin, 22307 LANNION, FRANCE

ABSTRACT

An adaptive filter in a blind equalization scheme has recently been proposed in order to reduce telephone line effects for speech recognizers. This paper presents the principles of this filter and describes the implementation of a circular-convolution frequency domain adaptive filter in the blind equalization scheme. The property of a constant long-term speech spectrum helps to compute the gradient used for the adaptation of the weights. However, using this property in a straightforward manner results in a crude implementation of this filter. Alternative computations of the standard stochastic gradient algorithm are therefore evaluated. On the basis of the speech recognition results obtained from different speaker independent telephone databases, this filter proves to be efficient for the channel equalization task.

1. INTRODUCTION

The speech signal to be processed by speech recognizers is often corrupted by several types of distortion. One of these is additive and corresponds to the ambient noise. A second type of distortion is convolutive and corresponds to the channel effect. These two types of distortion are responsible for the reduction of the discrimination capacity of the models and deteriorate the performances of speech recognizers.

This paper focuses on the convolved distortion problem and proposes a comprehensive evaluation of a new channel equalization technique [1]. Section 2 presents the principles of the blind adaptation scheme. The circular-convolution frequency domain adaptive filter used to implement this equalization technique is then described. Section 3 presents some implementation considerations (variants around LMS algorithms and variable reference spectrum). Finally, speech recognition performances, evaluated on several databases recorded over the telephone network, are provided in Section 4.

2. BLIND EQUALIZATION SCHEME

The block diagram of the blind equalization scheme is given in Figure 1.

The speaker produces a speech signal *s(t)* to which additive ambient noise *n(t)* is added. The resulting signal *x(t)* is captured by the handset microphone and is conveyed by the telephone line to the speech recognition system. Under the hypothesis of a convolved telephone line effect, the observed input signal *y(t)* may be written:

$$y(t) = [s(t)+n(t)] \otimes w(t) = x(t) \otimes w(t) \quad (1)$$

where $\otimes$ represents the convolution operator and *w(t)* is a linear and time invariant (LTI) filter.

The idea consists in defining an adaptive filter *h(t)* which unravels the line effect from the observed signal *y(t)* to produce $\hat{x}(t)$.

Spectral densities of the signal frames are available at the output of a Mel frequency filterbank in the speech recognition system. It is therefore easier to implement the filter in the frequency domain using a circular-convolution frequency domain adaptive filter implementation [2]. Equation 1 may be rewritten:

$$\Gamma_y(f) = \Gamma_x(f) W^2(f) \quad (2)$$

and

$$\Gamma_{\hat{x}}(f) = \Gamma_x(f) W^2(f) H^2(f) \quad (3)$$

where $\Gamma_{\hat{x}}(f)$ and $\Gamma_x(f)$ represent the spectral densities of $\hat{x}(t)$ and *x(t)*, and where *W(f)* and *H(f)* denote respectively the magnitude of the transfer functions of the telephone channel and the adaptive filter.

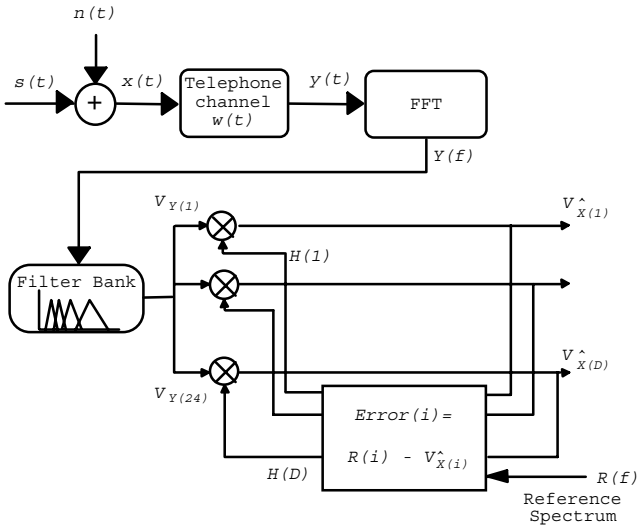

Figure 1. Blind Equalization Scheme.

The property of a constant long-term spectrum of speech (*R(f)*) is used to adapt the filter parameters.

The error can be written:

$$Error(f) = R(f) - \Gamma_{\hat{x}}(f) \quad (4)$$

The minimum mean-square error (MMSE) criterion determines $H_{opt}(f)$. The mean-square error can be written:

$$E\left(Error^2(f)\right) = E\left((R(f) - \Gamma_x(f)\, W^2(f)\, H^2(f))^2\right) \quad (5)$$

Minimizing this error yields the optimal filter:

$$H^2_{opt}(f) = \frac{R(f)\,\overline{\Gamma}_x(f)}{\overline{\Gamma_x^2}(f)} \frac{1}{W^2(f)} = cte\, \frac{1}{W^2(f)} \quad (6)$$

The optimal filter compensates the convolved effect of the telephone channel and enables the equalization of the channel effect.

3. IMPLEMENTATION CONSIDERATIONS

The LMS algorithm is the most popular technique for adaptive filtering applications and variants are proposed to improve this algorithm. Some variants around the LMS algorithm are evaluated in the context of this equalization scheme for speech recognition.

3.1. Variants around the LMS Algorithm

The standard LMS algorithm can be written:

$$H_{n+1}(i) = H_n(i) + \mu\, V_{Yn(i)}\, (R_n(i) - H_n(i)\, V_{Yn(i)}) \quad (7)$$

where $V_{Yn(i)}$ is the ith output of the MEL filterbank, $H(i)$ is the related weight, μ is the step size and the underscript n indicates the nth frame.

The convergence rate of a gradient-descent algorithm is determined by the eigenvalue disparity of the input signal correlation matrix. These eigenvalues correspond roughly to the power of the signal spectrum at equally-spaced frequency points around the unit circle [2]. Therefore this power variation can be compensated by using step sizes that are inversely proportional to the power levels in the frequency bins. Frequency domain filtering can therefore easily be modified to allow more uniform convergence by using the normalized LMS algorithm:

$$H_{n+1}(i) = H_n(i) + \mu\, (R_n(i) / V_{Yn(i)} - H_n(i)) \quad (8)$$

Signed algorithms are other alternatives to the standard time domain LMS algorithm [6]. The signed-regressor LMS algorithm (Eq. 9), the signed-error LMS algorithm (Eq.10) and the sign-sign LMS algorithm (Eq.11) are three variants applied to this frequency domain block LMS:

$$H_{n+1}(i) = H_n(i) + \mu\, sign(V_{Yn(i)})\, (R_n(i) - H_n(i)\, V_{yn(i)}) \quad (9)$$

$$H_{n+1}(i) = H_n(i) + \mu\, V_{Yn(i)}\, sign(R_n(i) - H_n(i)\, V_{yn(i)}) \quad (10)$$

$$H_{n+1}(i) = H_n(i) + \mu\, sign(V_{Yn(i)})\, sign(R_n(i) - H_n(i)\, V_{yn(i)}) \quad (11)$$

3.2. Variable Reference Spectrum

Speech signal parts are more appropriate in order to obtain an estimate of the channel effects [1]. The energy of the signal input can be efficiently used for speech detection [3]. The reference spectrum is therefore chosen to vary with the energy of the input signal, so that the filter is less active when there is a non-speech input signal.

3.3. Step Size Tuning

It is widely accepted that the step size controls the convergence rate and steady-state performances of gradient-based algorithms. In the special case of channel equalization for speech recognizers, the step size trade-off can be physically interpreted. As the adaptive scheme is based on long-term statistics, the step size must be high enough to guarantee a fast filter convergence in a long-term sense and low enough to avoid the disturbance of any short-term information related to the speech sounds.

4. RECOGNITION EXPERIMENTS

The equalization filter is intended to be a part of the front-end module of a speech recognition system. Variants around the LMS algorithm and the usefulness of variable reference spectrum are therefore evaluated on the basis of speech recognition experiments.

These recognition experiments are carried out with training and testing on equalized data. The first set of experiments aims at comparing the performances obtained with different variants around the LMS algorithm. The utility of a variable reference spectrum is examined. Then the proposed filter is compared to the IIR highpass filtering of cepstral trajectories and the cepstral subtraction technique [5]. Finally the compatibility of different equalization techniques with HMM models trained with non-equalized data is investigated.

4.1. Databases

The French speech databases used in these experiments were recorded using 800 speakers with different regional accents. The data was collected over the telephone network (mainly long distance calls). Utterances were automatically detected, keeping several silente frames on each side of the words. The detected words were then checked by native listeners. Each database is split into 2 parts, half is used to train the model parameters, and the other half is used to evaluate the recognition performances. The vocabulary of the DG1 database is the ten digits. The vocabulary of the NB2 database is 2-digit numbers (from 00 to 99) and the vocabulary of the TRG database is composed of 36 French command words. Table 1 gives the size of these speech databases.

Table 1. Size of Databases.

Database	DG1	NB2	TRG
Number of tokens	2 x 4700	2 x 7200	2 x 12200

4.2. Modelization

The speech recognition system used in this study is based on a Markov modeling approach. Every 16 ms, the acoustic analysis computes a set of 8 Mel frequency cepstral coefficients plus the energy of the frame. The 9 first order, and 9 second order temporal derivatives are then estimated using a frame window, which is centered on the current frame. Consequently the acoustical vectors have 27 components.

The speech recognition system uses a fully compiled network which includes a model for start and end silences and accepts various kinds of basic units (words, phonemes, allophones, etc.) [4]. Word models are used in this contribution.

4.3. Variants around the LMS Algorithm

Figure 2 shows speech recognition performances of the standard LMS algorithm. This algorithm was tested for a set of step size values. The performances obtained with the three best values of the step size are presented in figure 2. The performance of the baseline system is also presented, thus allowing the comparison of performances. Results show that no improvement of speech recognition performances is obtained with the standard LMS algorithm. With a low value of step size, the filter has insufficient time to converge and therefore does not perform any equalization. Thus no effect on the speech recognition performances can be observed with a low value of step size. Unfortunately, speech recognition performances deteriorate as the step size value increases.

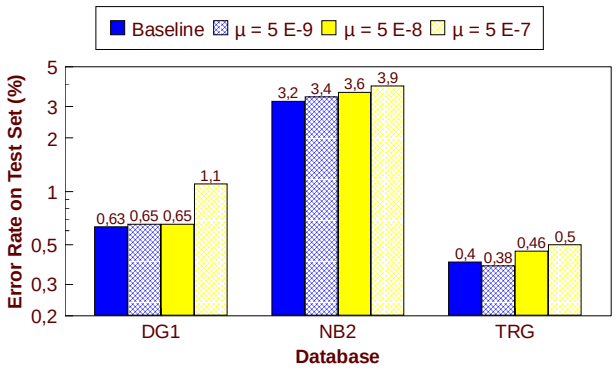

Figure 2. Standard LMS Algorithm.

Figure 3 shows speech recognition performances for the normalized LMS algorithm. No improvement is obtained for the DG1 database, while speech recognition performances are increased for the others databases. As expected, normalized LMS performs better than standard LMS.

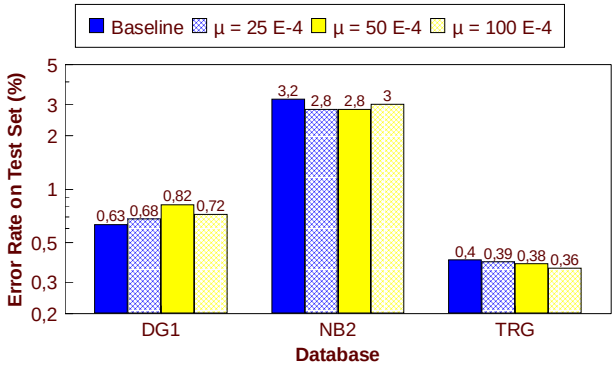

Figure 3. Normalized LMS Algorithm.

The performances of the three signed LMS algorithms, i.e. the signed-regressor LMS (SRLMS on the legend), the signed-error LMS (SELMS) and the sign-sign LMS (SSLMS), obtained with their best step size, are reported in figure 4. The performances of the baseline system and the normalized LMS algorithm (NLMS) are also presented.

Figure 4 shows that the signed-error LMS algorithm allows slight improvements on the DG1 and TRG databases but that performances deteriorate for the NB2 database. The sign-sign LMS algorithm allows slight improvement of speech recognition performances on the DG1 and NB2 databases. However, performances decrease on the TRG database. If we look at figure 4, we can see that the signed-regressor LMS algorithm is the only algorithm that improves speech recognition on the three databases. This algorithm outperforms the normalized LMS algorithm.

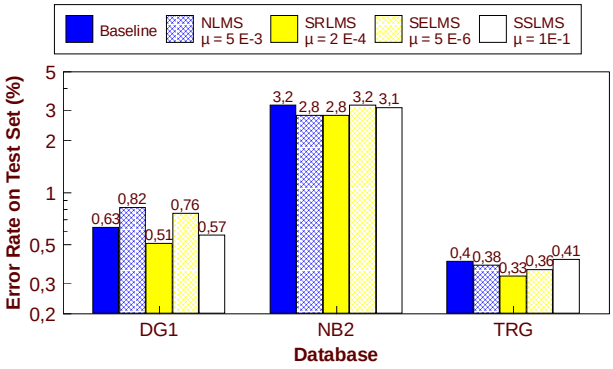

Figure 4. Signed LMS Algorithms.

4.4. Variable Reference Spectrum

The two best variants around the LMS algorithm, the normalized LMS with step size equal to 5 E-3, and the signed-regressor LMS with step size equal to 20 E-5, are stored to investigate the usefulness of a variable reference spectrum.

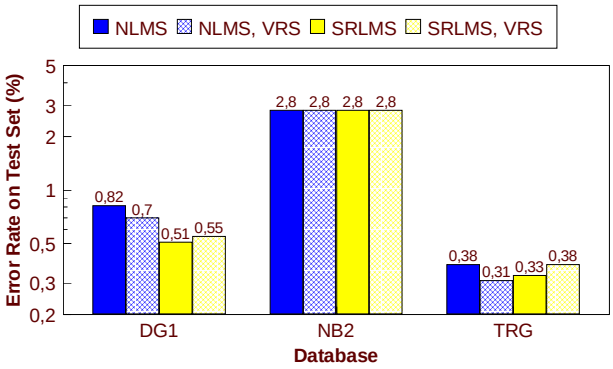

Figure 5. Utility of a Variable Reference Spectrum.

Figure 5 shows that a variable reference spectrum (VRS added to the legend) is useful for the normalized LMS. Recognition performances are improved on the DG1 and TRG databases and remain constant on the NB2 database. Conversely, for the signed-regressor LMS, recognition performances deteriorate on the DG1 and TRG databases and remain constant on the NB2 database. A variable reference spectrum is therefore damaging for the signed-regressor LMS algorithm.

4.5. Comparison of different techniques

In this experiment, the two best variants around the LMS algorithm for the blind equalization approach are compared to the highpass IIR filtering of cepstral trajectories and the cepstral subtraction technique. The normalized LMS is used with a variable reference spectrum as opposed to the signed-regressor LMS. The highpass filtering of cepstral trajectories (IIR on the legend) consists in applying a highpass filter to the cepstral trajectories. The cepstral subtraction technique (MCS on the legend) consists in estimating the long-term cepstra of speech for a given call and subtracting it from the cepstra of the frames. Results are shown in figure 6.

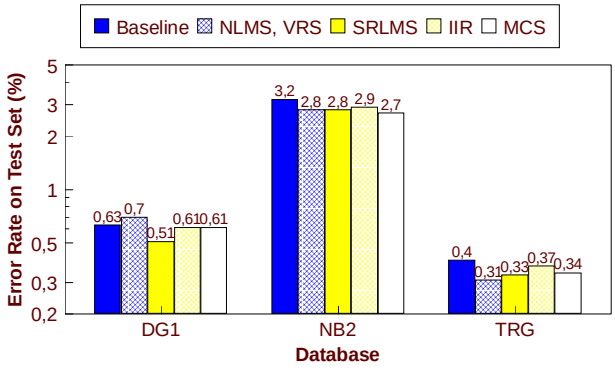

Figure 6. Different Channel Equalization Techniques.

It appears that the signed-regressor LMS algorithm is the only algorithm that outperforms the IIR highpass filtering on the three databases. The error rate reduction obtained on the three databases with the IIR highpass filtering, the cepstral subtraction and the signed-regressor LMS algorithm are reported in table 2.

Table 2. Error Rate Reduction.

Database	DG1	NB2	TRG
IIR highpass filtering	3 %	8,6 %	7,8 %
Cepstral subtraction	3 %	16,8 %	13,7 %
Signed-regressor LMS	20 %	12,9 %	17,6 %

If we look at table 2, we can see that the blind equalization approach outperforms the high-pass filtering technique, an on-line approach. It also gives slightly better or almost identical performances, depending on the database, as the cepstral subtraction technique, an off-line approach.

4.6. Compatibility with 'non-equalized' HMM models

An equalization process in the front-end of a speech recognition system enhances performances. An important factor is to know if it is necessary to retrain HMM models, or if equalization can also improve performances with models trained with non-equalized data.

In the final experiment, the blind equalization approach, the highpass IIR filtering and the cepstral subtraction technique are compared on the basis of speech recognition performances obtained with models trained with non-equalized data. Results are shown in figure 7. Speech recognition performances tend to decrease slightly when an equalization technique is used with models trained with non-equalized data.

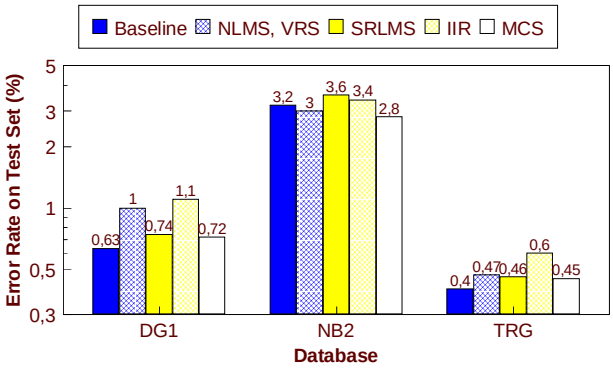

Figure 7. Compatibility with non-equalized HMM models.

5. CONCLUSION AND FUTURE WORK

This paper proposes a comprehensive evaluation of a frequency domain adaptive filter to implement a channel equalization technique for speech recognition. The signed-regressor LMS appears to be the best variant around the LMS algorithm. The proposed filter, an on-line technique, outperforms the IIR highpass filtering of cepstral trajectories, also an on-line approach, and provides slightly better performances than the cepstral subtraction technique which is an off-line approach.

Future studies will be devoted to the evaluation of the blind equalization technique on field databases. The rejection of out-of-vocabulary words is currently the main problem to be solved when building high performance speech recognition systems [7]. The effect of using different equalization techniques on the capability of the system to reject extraneous speech will also be studied. Finally, equalization of the energy parameter will be investigated, as well as the optimal association of this filter with a speech detector.

REFERENCES

[1] **C. Mokbel, D. Jouvet, J. Monné.** Blind Equalization using Adaptive Filtering for improving Speech Recognition over Telephone. Eurospeech 95, Madrid, pp 1987-1990, September 1995.

[2] **J.J. Shynk.** Frequency-Domain and Multirate Adaptive Filtering. IEEE SP Magazine, pp 15-37, January 1992.

[3] **L. Mauuary, J. Monné.** Speech/non-Speech Detection for Voice Response Systems. Eurospeech 93, Berlin, pp 1097-1100, September 1993.

[4] **D. Jouvet, K. Bartkova, J. Monné.** On the Modelization of Allophones in a HMM based Speech Recognition System. Eurospeech 91, Genova, pp 923-926, September 1991.

[5] **H. Murveit, J. Butzberger, M. Weintraub.** Reduced Channel Dependence for Speech Recognition. Proc. Speech and Natural Language Workshop, 1992, pp 280-284.

[6] **C. R. Johnson, Jr.** On the Interaction of Adaptive Filtering, Identification, and Control. IEEE Signal Processing Magazine, pp 22-37, March 1995.

[7] **L. Mauuary.** Improvements of the Performances of Interactive Voice Response Services. Doctoral Thesis, Université de Rennes, January 1994 (in French).