A MODIFIED EPHRAIM-MALAH NOISE SUPPRESSION RULE FOR AUTOMATIC SPEECH RECOGNITION

Roberto Gemello[*], *Franco Mana*[*] *and Renato De Mori*[§]

[*] LOQUENDO
via Nole, 55 – 10149 Torino - Italy
roberto.gemello@loquendo.com
franco.mana@loquendo.com

[§] LIA CNRS
University of Avignon , BP 1228
84911 Avignon Cedex 9 - France
renato.demori@lia.univ-avignon.fr

ABSTRACT

A soft decision gain modification is introduced and applied to the Ephraim-Malah gain function based on Maximum Mean Square Error Estimation (MMSE) [4-5] after amplitude compression. Non-linear evaluations of the noise overestimation factor and spectral floor are used in the same way for the proposed gain modification and for non-linear spectral subtraction (NSS). Consistent and statistically significant ASR improvements of the proposed approach with respect to NSS are observed for different noise conditions considered in the AURORA2 and AURORA3 corpora. As the non-linearity affects the two approaches in the same way, the result of comparison is particularly interesting.

1. INTRODUCTION

There is unquestionable evidence that additive noise, frequently present in many real-life situations, may strongly affect speech intelligibility and the performance of Automatic Speech Recognition (ASR) systems. Many solutions have been proposed for enhancing speech in order to make it more understandable and recognizable when it is corrupted by noise. Uncorrelated additive noise is frequent in many real-life situations and a great attention has been devoted to reduce the distortion introduced by this type of noise.

In the case of ASR, noise makes more severe the mismatch between training conditions, in which samples are collected for inferring the parameters of acoustic models, and test conditions. Essentially, two major approaches can be taken to reduce such a mismatch, namely, transforming the descriptors of the speech signal and adapting the models. Parameter transformation can be based on a theory that does not require any training or on functions whose parameters have to be inferred by the statistical analysis of a training corpus.

Different approaches can be combined and certain combinations may lead to improvements with respect to the use of a single approach.

This paper focuses on the use, in ASR, of gain functions that multiply noisy acoustic parameters transforming them into estimations of clean speech parameters, without any training involving a specific corpus. Among other possibilities, a gain can be expressed by the magnitude of the transfer function of a Wiener filter that attempts to subtract the noise component from the spectrum of a noisy speech signal. Recently, attempts have been made to incorporate some perceptual findings into this type of spectral subtraction. In [9], a non-linear spectral subtraction is proposed, motivated by the fact that, for spectral peacks, the signal has enough energy to mask the residual noise. Thus, for a specific frequency bin, the residual noise will not be perceived. This is not the case for spectral valleys where a residual noise can be perceived. Moreover, human perception is less sensitive to spectral valleys suggesting to overestimate the noise component in these regions and perform a Nonlinear Spectral Subtraction (NSS). NSS appears to be beneficial not only for speech coding and transmission, but also for ASR [2], [7], [10]. The explanation could be that, in both cases, a non-linear compression of spectral samples is performed in such a way that the effects of noise do not perturb too much the spectral samples corresponding to peaks of the speech component; while the samples of spectral valleys, are strongly attenuated.

Unfortunately, even the application of non-linear techniques may leave residual distortions and it is interesting to investigate with which approach these distortions introduce less damage for ASR.

A soft-decision gain modification for speech enhancement (but not for speech recognition) has been proposed in [8] and modified in [3] with the introduction of the a-priori speech absence probability (SAP). SAP is computed for each frequency bin using a global frame probability evaluated with heuristic considerations. In this paper, a different soft decision gain modification is introduced and applied to the Ephraim-Malah gain function based on Maximum Mean Square Error Estimation (MMSE) [4-5] after amplitude compression. Non-linear evaluations of the noise overestimation factor and spectral floor are used in the same way for the proposed gain modification and for NSS with Wiener filter. Consistent and statistically significant ASR improvements of the proposed approach with respect to NSS are observed for different noise conditions considered in the AURORA2 and AURORA3 corpora. As the non-linearity affects the two approaches in the same way, the result of comparison is particularly interesting.

Basic theory and proposed modifications are described in sections 2 and 3, while experimental set up and results obtained with the AURORA2 and AURORA3 corpora are described in section 4. The main focus of the paper being denoising, the ASR system was not trained nor adapted to the domain and the types of noise of the corpora.

2. BACKGROUND

Let {y(nT)} be a sequence of samples of a noisy speech signal; T is the time sampling period and n is the time sample index. Let {x(nT)} be the sequence of samples of the corresponding clean speech signal and {d(nT)} be a sequence of samples of additive noise which is uncorrelated with the clean speech. This is a frequent situation real-life ASR systems have to deal with.

Let $|Y_k(m)|^2$ be the k-th frequency sample of the spectrum energy of {y(nT)}, computed in the m-th time window. Let $|X_k(m)|^2$ and $|D_k(m)|^2$ be the k-th spectrum energy sample, computed in the m-th time window, of the clean signal and the noise, respectively. In order to adapt test conditions, in which noisy signals have to be recognized, to train conditions in which clean signals have been used, algorithms have been proposed for estimating $|X_k(m)|^2$ from the observation of $|Y_k(m)|^2$. A popular algorithm for this purpose uses a Wiener filter, whose transfer function is $G_k(m)$, to compute:

$$|\hat{X}_k(m)|^2 = G_k(m)\,|Y_k(m)|^2 \qquad (1)$$

It has been found [9] that better recognition performance is obtained if the transfer function is conceived to perform a non-linear spectral subtraction. In [7] and [10] it has been found that good results are obtained if the filter is used to perform a non-linear spectral subtraction as follows:

$$|X_k(m)|^2 = \begin{cases} \dfrac{\left[|Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2\right]^2}{|Y_k(m)|^2} & \text{if } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{otherwise} \end{cases} \qquad (2)$$

where $\alpha(m)$ is a noise overestimation factor, and $\beta(m)$ is a spectral floor used to avoid negative spectrum values. These two parameters vary in time as function of the Signal-to-Noise Ratio SNR(m), computed as follows:

$$SNR(m) = 10\log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right) \qquad (3)$$

where $|\hat{D}_k(m)|^2$ is an estimation of the k-th noise spectral sample at time m; $\alpha(m)$ and $\beta(m)$ can be defined as possibility functions of SNR(m) as shown in Figure 1.

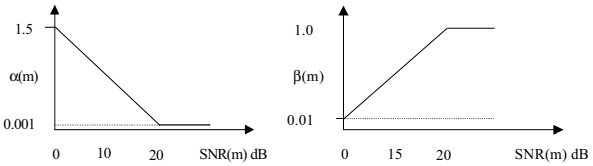

Figure 1 – Examples of definitions of $\alpha(m)$ and $\beta(m)$ as functions of SNR(m)

$G_k(m)$ can also be obtained with an approach proposed by (Ephraim and Malah) in [4-5]. In particular Ephraim–Malah MMSE log estimator is a short-time spectral amplitude estimator that minimizes the mean-square error of the estimated logarithms of the spectra, and it is well known that a distortion measure which operates on these logarithms is more suitable for speech processing than measures taken on the power spectra. It is defined as follows:

$$G_k = \frac{\xi_k(m)}{1+\xi_k(m)} \exp\left(\frac{1}{2}\int_{\nu_k(m)}^{\infty} \frac{e^{-t}}{t}\,dt\right) \qquad (4)$$

where:

$$\xi_k(m) = \frac{|X_k(m)|^2}{|D_k(m)|^2} \quad \text{is the } \textit{a priori} \text{ SNR,} \qquad (5)$$

$$\gamma_k(m) = \frac{|Y_k(m)|^2}{|D_k(m)|^2} \quad \text{is the } \textit{a posteriori} \text{ SNR,} \qquad (6)$$

and $\nu_k(m) = \dfrac{\xi_k(m)}{1+\xi_k(m)}\gamma_k(m)$,

The computation of the *a priori* SNR requires the knowledge of the clean speech spectrum, which is not available. An estimation can be obtained with *a decision-directed approach* [4] as follows:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + [1-\eta(m)]\max[0, \gamma_k(m)-1], \quad \eta(m)\in[0,1) \qquad (7)$$

In [1], it is shown that it is convenient for speech coding to make $\eta(m)$ dependent on the global SNR(m) and to assign to it a high value if SNR(m) is low and a low value if SNR(m) is high. This variation is adopted in our experimentation.

3. PROPOSED METHOD

In this paper we propose a modification of Ephraim-Malah gain by making the estimation of the *a priori* and the *a posteriori* SNR dependent on the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ as follows:

$$\tilde{\hat{\xi}}_k(m) = \max\left(\eta\frac{|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta)[\tilde{\gamma}_k(m)-1], \beta(m)\right), \quad \eta(m)\in[0,1) \qquad (8)$$

$$\tilde{\gamma}_k(m) = \max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\right) + 1 \qquad (9)$$

where the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ vary with global SNR(m) as shown in Figure 1. The adopted approach modifies the estimates of γ_k and ξ_k while maintaining the global shape of the gain function $G_k(\gamma_k, \xi_k)$. The modified gain function can be expressed as follows:

$$\tilde{G}_k(\gamma_k(m), \xi_k(m)) = G_k\left(\tilde{\gamma}_k(m), \tilde{\xi}_k(m)\right)$$

with $\tilde{\xi}_k(m), \tilde{\gamma}_k(m)$ computed according to (8) and (9).

Figure 2 shows the original Ephraim-Malah attenuation rule (4) and its version with the above-proposed modifications.

The effect of the introduced modification is a gradual reduction of the attenuation produced by the original gain in areas of high posterior SNR γ_k, as the global SNR increases.

Noise estimation appears in the computation of (8) and (9). For the experiments described in this paper, an estimate of the noise spectrum amplitude is obtained by a first-order recursion in conjunction with an energy based Voice Activity Detector (VAD) as follows [7]:

$$\hat{D}_k(m) = \begin{cases} \lambda\hat{D}_k(m-1) + (1-\lambda)Y_k(m) \\ \quad \textit{if} \quad \left\{ |Y_k(m) - \hat{D}_k(m)| \le \mu\sigma(m) \right\} \wedge \{VAD = false\} \\ \hat{D}_k(m-1) \qquad \textit{otherwise} \end{cases} \qquad (10)$$

where, λ controls the update speed of the recursion and μ the allowed dynamics of noise; σ(m) is the noise standard deviation, estimated as:

$$\sigma^2(m) = \gamma\sigma^2(m-1) + (1-\gamma)\left(Y_k(m) - \hat{D}_k(m)\right)^2 \qquad (11)$$

The values for λ and μ *are* respectively 0.9 and 4.0.

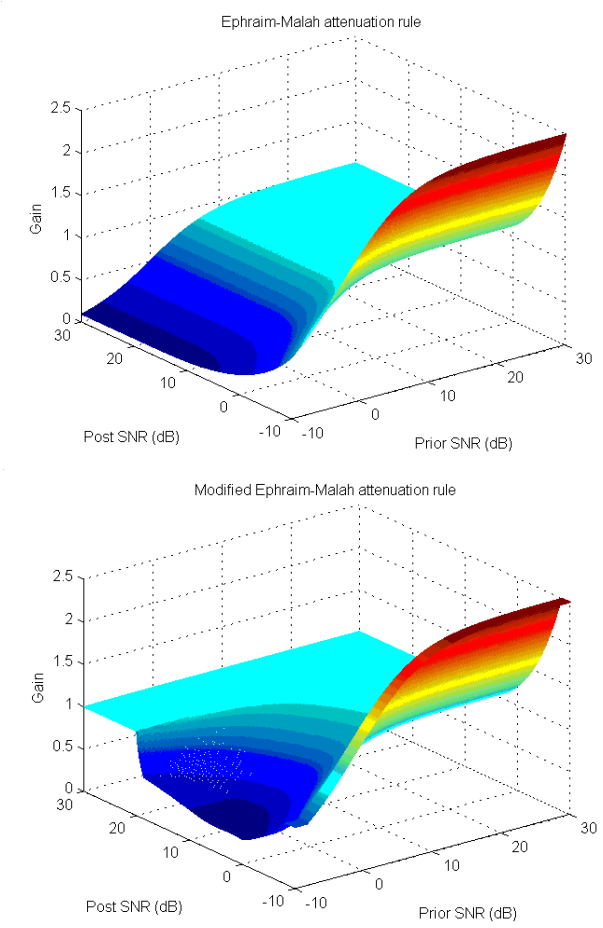

Figure 2 - Original attenuation rule computed with (4) and its version with the proposed modifications (8) and (9).

4. EXPERIMENTAL SETUP AND RESULTS

Experiments were conducted with a hybrid HMM-NN ASR system described in [6]. As the purpose of this paper is to compare different denoising methods whose parameters are independent from the recognizer, the results should not depend on the type of recognizer used. The ASR has been trained for the target languages using large, domain and task independent corpora, like SpeechDat1-2, not collected in noisy environments and without added noise. Aurora2-3 corpora were not used to train the ASRs but just to test them. Acoustic modeling was made using phonetical sub-word units instead of whole word models.

Aurora2 and Aurora3 data were used for comparing the performance of baseline Wiener filtering, SNR-dependent Wiener filtering, baseline Ephraim-Malah short-time spectral attenuation rule based on log estimator and the proposed modification of the Ephraim-Malah rule. Tables I-VI show the results on Aurora2 expressed in Word Accuracy (%). Confidence interval on WA is 0.2%. Averages are computed in the 0-20 dB range.

Experimental results show that:

- Ephraim-Malah gain outperforms Wiener gain in its baseline version;
- this tendency is confirmed when using the modified version of the rules as proposed in [7] for Wiener gain and in this paper for Ephraim-Malah gain;
- The best results are always obtained with the modified Ephraim-Malah gain, with the only exception of test C (mainly conceived for testing channel mismatch) where the best result is obtained with baseline Ephraim-Malah gain.

Another test was performed on the Aurora3 corpus with Speech-Dat car connected digits in Italian, Spanish and German on the High Mismatch test set and on the noisy component (CH1) of the training set (used as test set)

.

Table I Test set A results for Aurora2 without denoising

SNR/dB	Subway	Babble	Car	Exhibit	Average
clean	99.4	98.9	99.0	99.0	**99.1**
20	98.6	98.6	98.5	98.5	**98.5**
15	97.1	97.1	97.7	96.0	**97.0**
10	91.7	92.3	91.9	89.0	**91.2**
5	76.6	78.0	76.4	70.4	**75.3**
0	47.0	51.9	34.6	36.1	**42.4**
-5	15.6	22.9	9.1	11.7	**14.8**
Average	**82.2**	**83.6**	**79.8**	**78.0**	**80.9**

Table II Test set A results for Aurora2. Denoising with baseline Wiener filtering

SNR/dB	Subway	Babble	Car	Exhibit	Average
clean	99.4	98.9	99.0	99.0	**99.1**
20	98.5	98.7	98.9	98.7	**98.7**
15	98.1	97.8	98.6	97.5	**98.0**
10	94.0	94.8	96.5	93.2	**94.6**
5	85.7	83.3	89.4	82.0	**85.1**
0	64.9	56.6	69.5	60.5	**62.9**
-5	35.8	25.0	31.3	30.3	**30.6**
Average	**88.2**	**86.2**	**90.6**	**86.4**	**87.9**

Table III Test set A results for Aurora2. Denoising with SNR-dependent Wiener filtering

SNR/dB	Subway	Babble	Car	Exhibit	Average
clean	99.4	98.9	99.1	99.0	**99.1**
20	98.7	98.8	98.9	98.8	**98.8**
15	98.3	98.0	98.5	97.5	**98.1**
10	93.7	95.0	96.5	93.5	**94.7**
5	85.1	84.5	90.0	82.0	**85.4**
0	63.9	60.0	70.9	60.2	**63.8**
-5	34.2	27.4	33.8	32.2	**31.9**
Average	**87.9**	**87.3**	**91.0**	**86.4**	**88.1**

Table IV Test set A results for Aurora2. Baseline Ephraim-Malah denoising

SNR/dB	Subway	Babble	Car	Exhibit	Average
clean	99.4	98.9	99.1	99.0	**99.1**
20	98.5	98.6	98.8	98.5	**98.6**
15	98.1	97.7	98.6	97.1	**97.9**
10	93.9	94.1	97.0	92.9	**94.5**
5	85.5	82.9	91.9	83.8	**85.8**
0	67.5	57.3	75.1	61.9	**65.4**
-5	38.2	25.9	40.0	35.2	**34.8**
Average	**88.7**	**86.1**	**92.1**	**86.8**	**88.4**

Table V Test set A results for Aurora2. Modified Ephraim-Malah denoising

SNR/dB	Subway	Babble	Car	Exhibit	Average
clean	99.4	98.9	99.0	99.0	**99.1**
20	98.6	98.7	98.8	98.9	**98.8**
15	98.3	98.0	98.7	97.4	**98.1**
10	94.3	94.6	97.2	93.7	**94.9**
5	85.5	84.2	91.9	82.9	**86.1**
0	66.7	61.4	76.1	63.5	**66.9**
-5	37.3	28.4	41.0	35.5	**35.5**
Average	**88.7**	**87.4**	**92.5**	**87.3**	**89.0**

Table VI Comparative denoising results for Aurora2 test sets A, B, C

Denoising Method	Test Set A	Test Set B	Test Set C
No Denoising	**80.9**	**83.2**	**77.5**
Wiener baseline	**87.9**	**88.1**	**86.4**
Wiener SNR dep	**88.1**	**88.3**	**86.3**
Ephraim-Malah baseline	**88.4**	**88.1**	**87.3**
Ephraim-Malah modif.	**89.0**	**88.6**	**86.9**

The tables VII and VIII compare the baseline of the two denoising gains (Wiener and Ephraim-Malah) and their modified versions on Aurora3 (expressed in WA %). Confidence interval for statistical relevance of results is shown for each test set.
Ephraim-Malah log estimator is always better than Wiener subtraction, both in the baseline version and in the modified version. The Ephraim-Malah modified gain obtains the best results, with an average error reduction of 8.4% w.r.t the Wiener SNR dep. gain, and an average 50.3% error reduction w.r.t. the case without denoising. The modification introduced in the Ephraim-Malah gain produces an average error reduction of 22.9% w.r.t. the baseline Ephraim-Malah gain.

Table VII Test on Aurora 3 High Mismatched test set

Aurora 3 Test on High Mismatched test set				
Denoising Method	Italian C.I. 1.4	Spanish C.I. 1.2	German C.I. 1.7	**Average**
No denoising	56.7	69.9	82.5	**69.7**
Wiener baseline	68.1	81.3	87.8	**79.1**
Wiener SNR dep.	74.9	86.2	89.2	**83.4**
Ephraim-Malah baseline	69.7	81.3	89.7	**80.2**
Ephraim-Malah modified	75.6	87.7	90.5	**84.6**

Table VIII Test on Aurora 3 Noisy component (CH1) of the train set (used as test)

Aurora 3 Test on CH1 component of train set				
Denoising Method	Italian C.I. 0.9	Spanish C.I. 0.6	German C.I. 1.0	**Average**
No denoising	58.6	73.8	85.8	**72.7**
Wiener baseline	70.7	81.3	89.9	**80.6**
Wiener SNR dep.	76.3	88.9	90.8	**85.3**
Ephraim-Malah baseline	71.5	85.9	90.6	**82.6**
Ephraim-Malah modified	77.5	90.6	92.1	**86.7**

5. CONCLUSIONS

In this paper, the application of Ephraim-Malah short-time spectral amplitude log estimator to speech recognition has been investigated. While widely and successfully used in speech enhancement, it is reported in the literature [11] that the application of the corresponding suppression rule does not result in a clear advantage over spectral subtraction when used in ASR.
The experiments described in this paper made evident that non-linear versions of these rules depending on global SNR reduce the WER in ASR when additive noise is present. Furthermore, the use of the non-linear technique proposed provides consistently better results when applied to the Ephraim-Malah attenuation rule based on MMSE log estimator with respect to the application to Wiener filtering.

6. REFERENCES

[1] C. Beaugeant, P. Scalart, Noise Reduction using Perceptual Spectral Change, *Eurospeech 1999*.

[2] J. Beh and H. Co, A novel spectral subtraction scheme for robust speech recognition : spectral subtraction using spectral harmonics of speech.*Proc. International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, pp. I-684-687,2003.

[3] I. Cohen ," Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator". *IEEE Signal Processing Letters*, 9,(4):11-117, 2002

[4] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimator", *IEEE Trans. Acoust. Speech Signal Processing*, vol ASSP-32, no. 6, pp. 1109-1121, 1984

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum min-square error log-spectral amplitude estimator", *IEEE Trans. Acoust. Speech Signal Processing*, vol ASSP-33, no. 2, pp. 443-445, 1985

[6] R. Gemello, D. Albesano, F. Mana, "Multi-source neural networks for speech recognition", in *Proc. of International Joint Conference on Neural Networks (IJCNN'99),* Washington, July 1999.

[7] R. Gemello, F. Mana, D. Albesano and R. De Mori, "Robust Multiple Resolution Analysis for Automatic Speech Recognition", *Eurospeech 2003*, Geneva, Switzerland.

[8] N.S. Kim and J.H. Chang, Spectral enhancement based on global soft decision. *IEEE Signal Processing Letters*, 7(5):108-110, 2000.

[9] P. Loockwood, J. Boundy, "Experiments with non-linear Spectral Subtractor (NSS), Hidden Markov Models, and the projection for robust speech recognition in cars", *Speech Communication* 11 (1992) 215-228.

[10] V. Schless, F. Class, SNR-Dependent flooring and noise overestimation for joint application of spectral subtraction and model combination, *ICSLP* 1998.

[11] M. Matassoni, G.A. Mian, M. Omologo, A, Santarelli and P. Svaizer, "Some experiments on the use of One-channel Noise Reduction techniques with the Italian Speechdat Car database", IEEE ASRU 2001, Madonna di Campiglio, Italy, December 2001.