Automatic Speech Recognition With a Modified Ephraim–Malah Rule

Article *in* Signal Processing Letters, IEEE · February 2006

DOI: 10.1109/LSP.2005.860535 · Source: IEEE Xplore

CITATIONS
19

READS
182

3 authors:

Roberto Gemello
Nuance Communications
86 PUBLICATIONS **976** CITATIONS
SEE PROFILE

Franco Mana
Nuance Communications
63 PUBLICATIONS **811** CITATIONS
SEE PROFILE

Renato De Mori
McGill University and University of Avignon
340 PUBLICATIONS **6,332** CITATIONS
SEE PROFILE

Automatic Speech Recognition with a Modified Ephraim-Malah Rule

Roberto Gemello, *Member, IEEE*, Franco Mana and Renato De Mori, *Fellow, IEEE*

***Abstract*—A soft decision gain modification is introduced and applied to the Ephraim-Malah gain function based on Maximum Mean Square Error Estimation (MMSE) 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 Aurora-2 and Aurora-3 corpora. As the non-linearity affects the two approaches in the same way, the result of comparison is particularly interesting.**

***Index Terms*— Speech recognition, Speech enhancement.**

I. 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 letter 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. The objective is to find a denoising procedure leading to an ASR performance close to the one observed when the ASR models are trained with a corpus collected with the lexicon and environment of the test 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 picks, 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 letter, 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 Aurora-2 and Aurora-3 corpora. As the non-linearity affects the two approaches in the same way, the result of comparison is particularly interesting.

Manuscript received May 11, 2005. This work was supported in part by the EC IST Project HIWIRE.

Roberto Gemello is with Loquendo, via Valdellatorre, 4, 10148 Torino, ITALY (e-mail: roberto.gemello@loquendo.com).

Franco Mana is with Loquendo, via Valdellatorre, 4, 10148 Torino, ITALY (e-mail: franco.mana@loquendo.com).

Renato De Mori is with University of Avignon, BP 1228, 84911 Avignon Cedex 9 – FRANCE (e-mail: renato.demori@lia.univ-avignon.fr).

Basic theory is briefly summarized in section II, while experimental set up and results obtained with the Aurora-2 and Aurora-3 corpora are described in section III. The main focus of the letter being denoising, the ASR system was not trained nor adapted to the domain and the types of noise of the corpora.

II. Background and Proposed Method

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:

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

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:

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

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:

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

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

$G_k(m)$ can also be obtained with an approach proposed 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:

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

where:

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

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

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:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + [1-\eta(m)]\max[0, \gamma_k(m)-1] \quad \eta(m) \in [0,1) \quad (8)$$

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

where the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ varies with SNR(m) as shown in Figure 1. Figure 2 shows a plot of the rule (4) and its versions with the above-proposed modifications at different levels of global SNR. 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:

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

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

Noise estimation appears in the computation of (8) and (9). For the experiments described in this letter, 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]:

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

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

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

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

III. Experimental Setup and Results

Experiments were conducted with a hybrid HMM-NN ASR system described in [6], consisting of a Neural Netwotk (NN) which computes observations for a set of Hidden Markov Models (HMM).

The testing conditions used in the experiments are the following:

- **No Denoising**: Rasta PLP features (RPLP) are used without any preliminary noise reduction.
- **Wiener baseline**: RPLP with denoising based on standard Wiener filtering.
- **Wiener modified**: RPLP with Wiener filtering dependent on global SNR [eq. (2)].
- **Ephraim-Malah baseline**: RPLP with denoising based on the standard Ephraim-Malah spectral attenuation rule [eq. (4) (5) (6)(7)].
- **Ephraim-Malah modified**: RPLP with denoising based on the modified Ephraim-Malah spectral attenuation rule [eq. (4) (8) (9)].
- **ETSI AFE**: The standard noise robust Advanced Front-End released by ETSI [12].

The first experiments were performed using the standard Aurora-2 speech corpus [13], that is made by the TI connected digits, downsampled to 8 kHz and with added noise. A NN has been training for each denoising condition on the two Aurora-2 train sets (Clean and Multi-condition). Acoustic modeling was made using phonetical sub-word units instead of whole word models.

Performance comparisons are shown in Table I. Results are expressed in percentage of Word Error Rate. Confidence interval on WA is 0.2%. Averages are computed in the 0-20 dB range, according to Aurora-2 reporting standard.

It is now interesting to compare these results with the ones obtained with recognition models trained using the large, domain independent, Macrophone American telephone corpus. Results are shown in Table II.

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 letter 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.

The second experiment was performed on the Aurora-3 corpus (connected digits recorded in car environment) in Italian, Spanish and German, on the High Mismatch test set and on the noisy component (CH1) of the training set (used only as test). The models have been trained with large, domain independent, telephone corpora, and Aurora-3 has been used only for test.

The tables III and IV report the results for the different conditions. Confidence intervals for statistical relevance of results are 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% with respect to 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% with respect to the baseline version.

IV. Conclusion

In this letter, 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 letter made evident that non-linear versions of these rules depending on global SNR effectively reduce the WER in ASR when additive noise is present. Furthermore, the use of non-linear techniques 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.

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 Workshop on Automatic Speech Recognition and Understanding (ASRU 2001), Madonna di Campiglio, Italia.

[12] ETSI ES 202 211 v1.1.1 2003-08, "Distributed Speech Recognition: Extended Front-End"

[13] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCAITRW ASR2000 "Automatic Speech*

Recognition: Challenge for the Next Millennium", Paris, France, September 2000

Roberto Gemello graduated in Computer Science at the University of Turin, Italy, in 1981. He is Senior Technical Leader in the Speech Technology Division of Loquendo (www.loquendo.com). Since 1986 he has been involved in automated learning systems, first in inductive symbolic systems and then in neural networks for speech recognition. He published more than 50 papers in international conferences and journals. He is co-inventor of nine international patents on neural networks, speech recognition and wavelets technology.

Franco Mana Born in 1963, Franco Mana graduated in Computer Science at the University of Turin, Italy, in 1987. He is Technical Leader in the Speech Technology Division of Loquendo. His main research field are neural networks applied to speech recognition. He published several papers in international conferences and journals and is co-inventor of three international patents on neural networks and wavelets technology.

Renato De Mori is professor of Computer Science at Mc Gill University (Canada) and at the University of Avignon (France). He is member of the IEEE Speech Technical Committee and Chief Editor of Speech Comunication. He serves in many Advisory boards including the Canadian Chairs Interdisciplinary Adjudication Committee. He is author of many papers on Computer Arithmetic, Sofware Engineering, Speech Processing and Recignition.

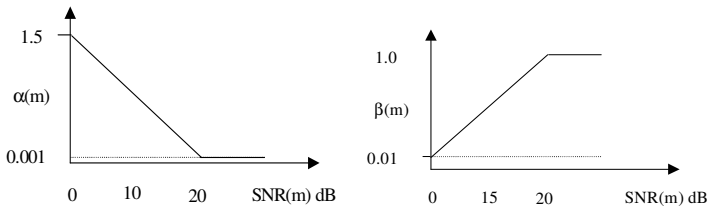
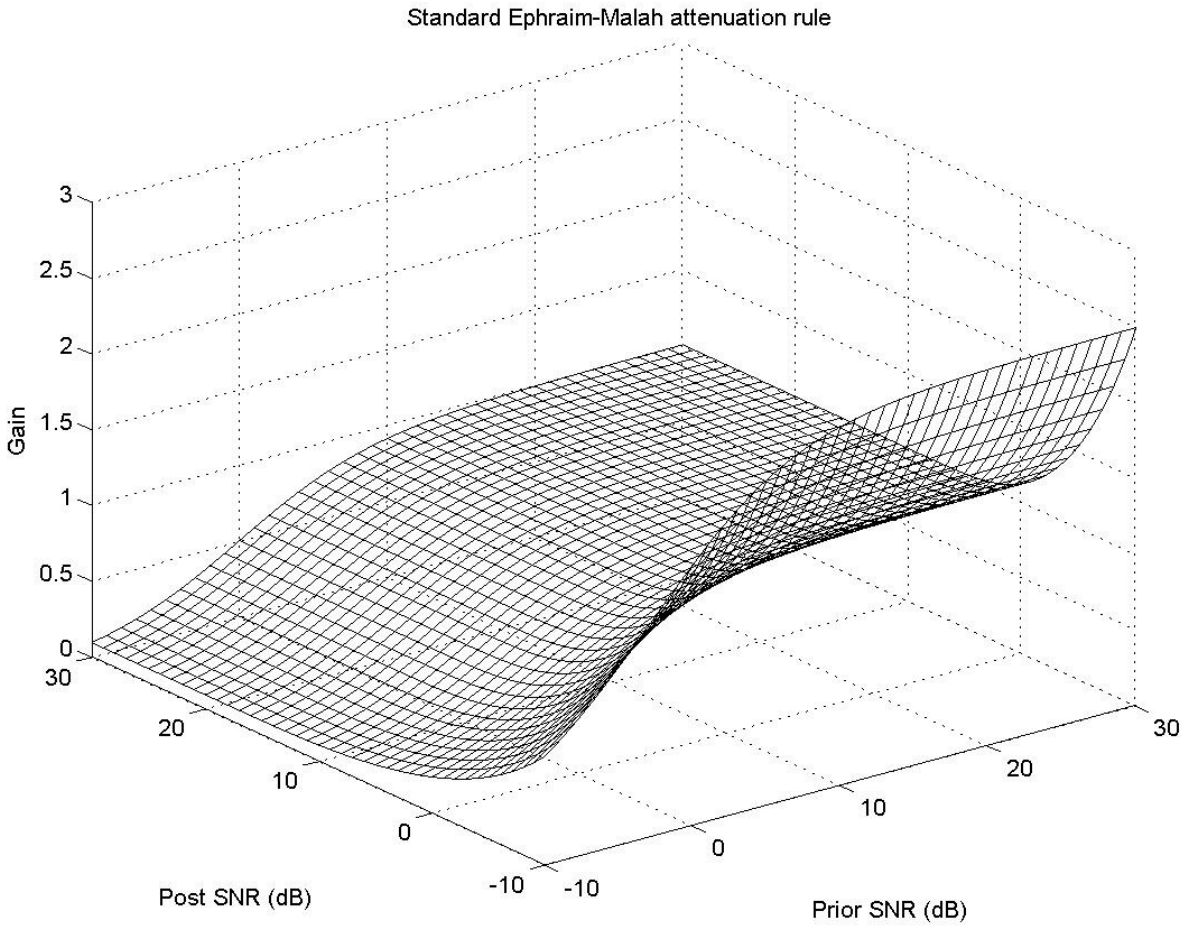
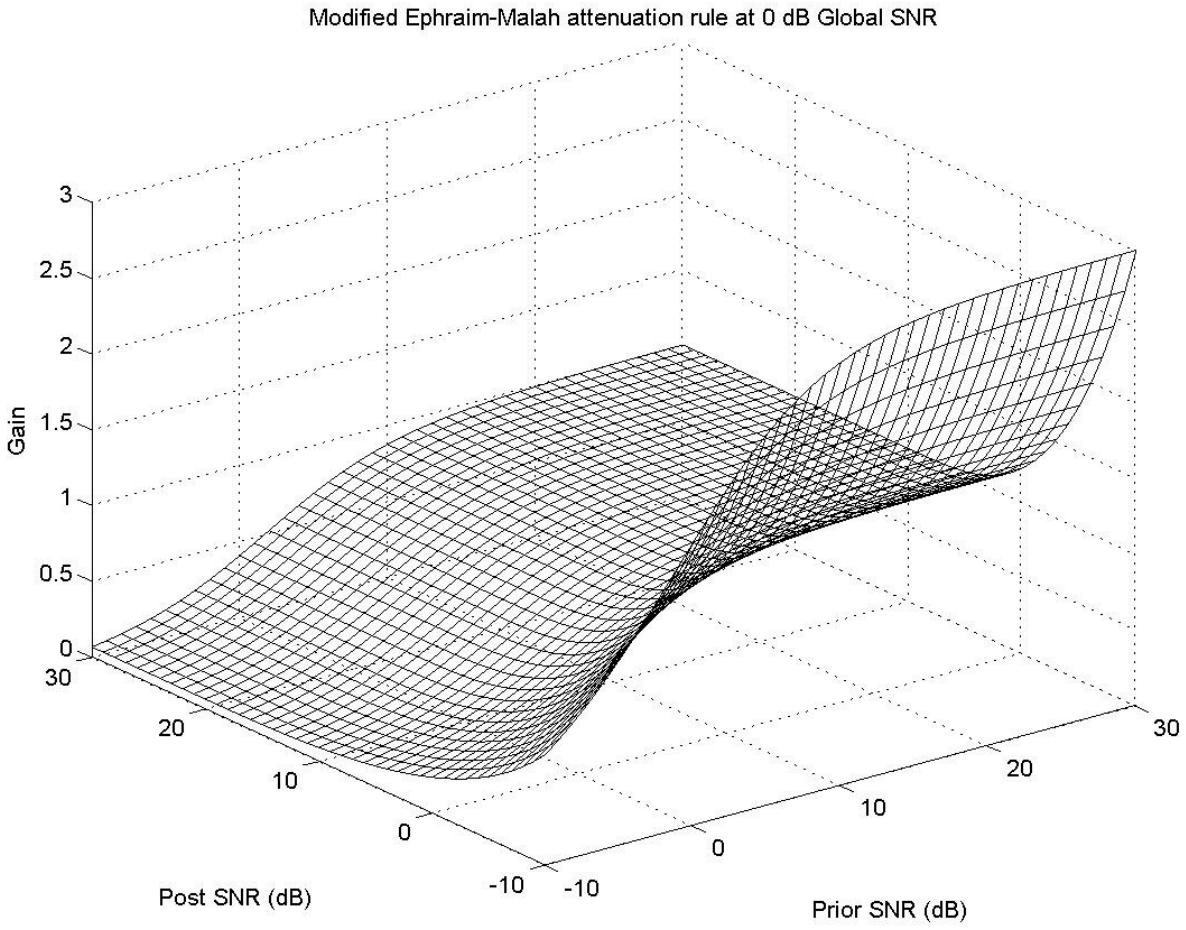
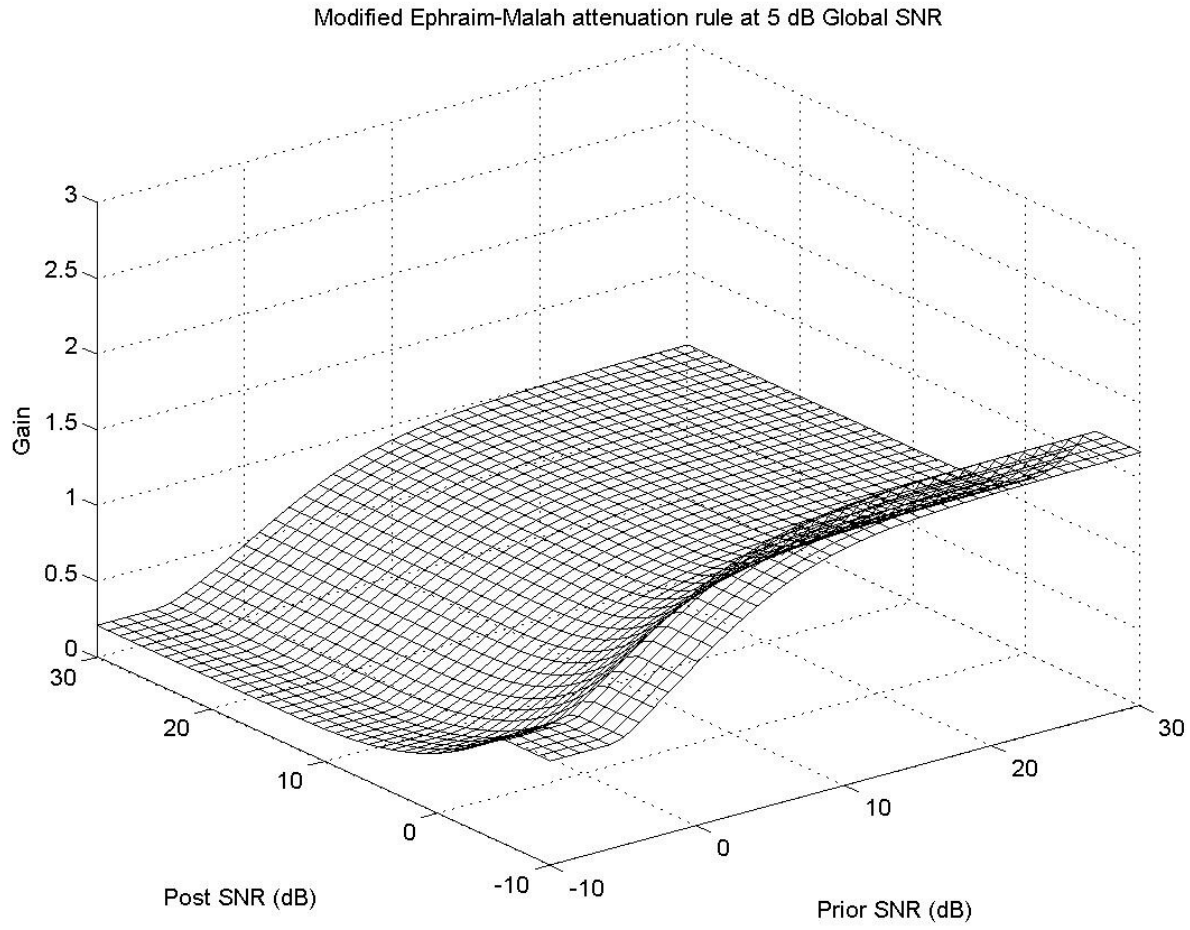
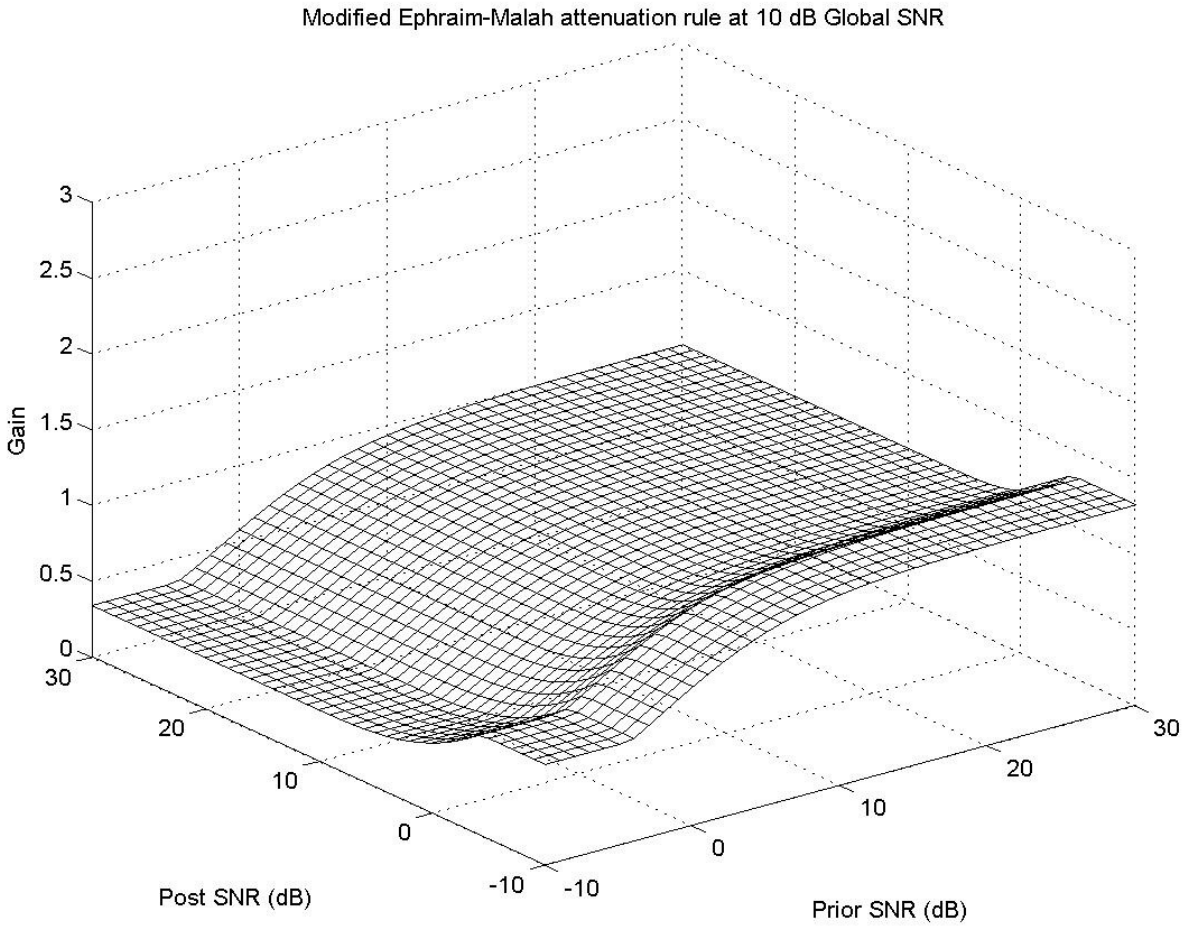


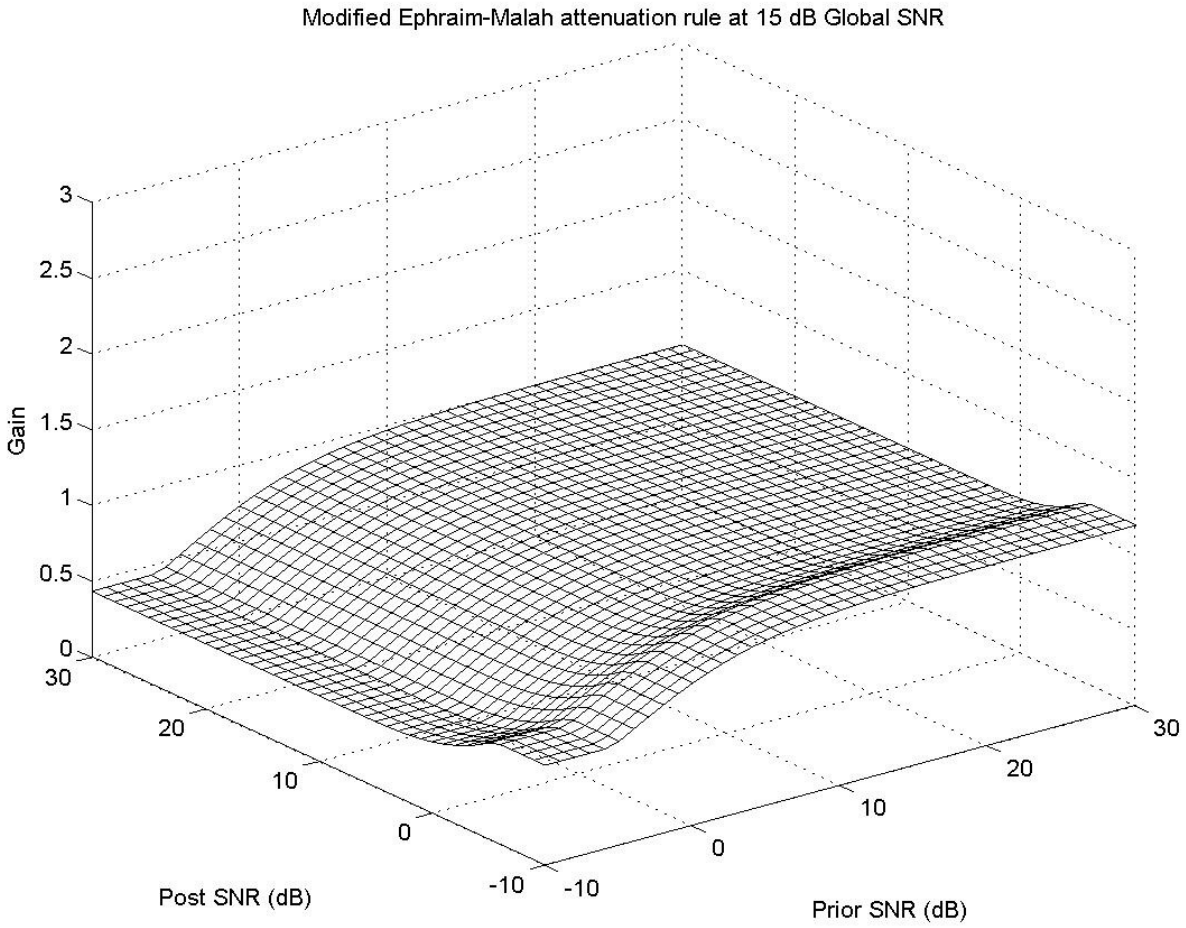
Fig. 1. Definition of $\alpha(m)$ and $\beta(m)$ as functions of SNR(m)


Standard Ephraim-Malah attenuation rule
Gain
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
Post SNR (dB)
-10
0
10
20
30
Prior SNR (dB)


Modified Ephraim-Malah attenuation rule at 0 dB Global SNR
Gain
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
Post SNR (dB)
-10
0
10
20
30
Prior SNR (dB)


Modified Ephraim-Malah attenuation rule at 5 dB Global SNR
Gain
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
Post SNR (dB)
-10
0
10
20
30
Prior SNR (dB)


Modified Ephraim-Malah attenuation rule at 10 dB Global SNR
Gain
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
Post SNR (dB)
-10
0
10
20
30
Prior SNR (dB)


Modified Ephraim-Malah attenuation rule at 15 dB Global SNR
Gain
3
2.5
2
1.5
1
0.5
0
30
20
10
0
-10
Post SNR (dB)
-10
0
10
20
30
Prior SNR (dB)

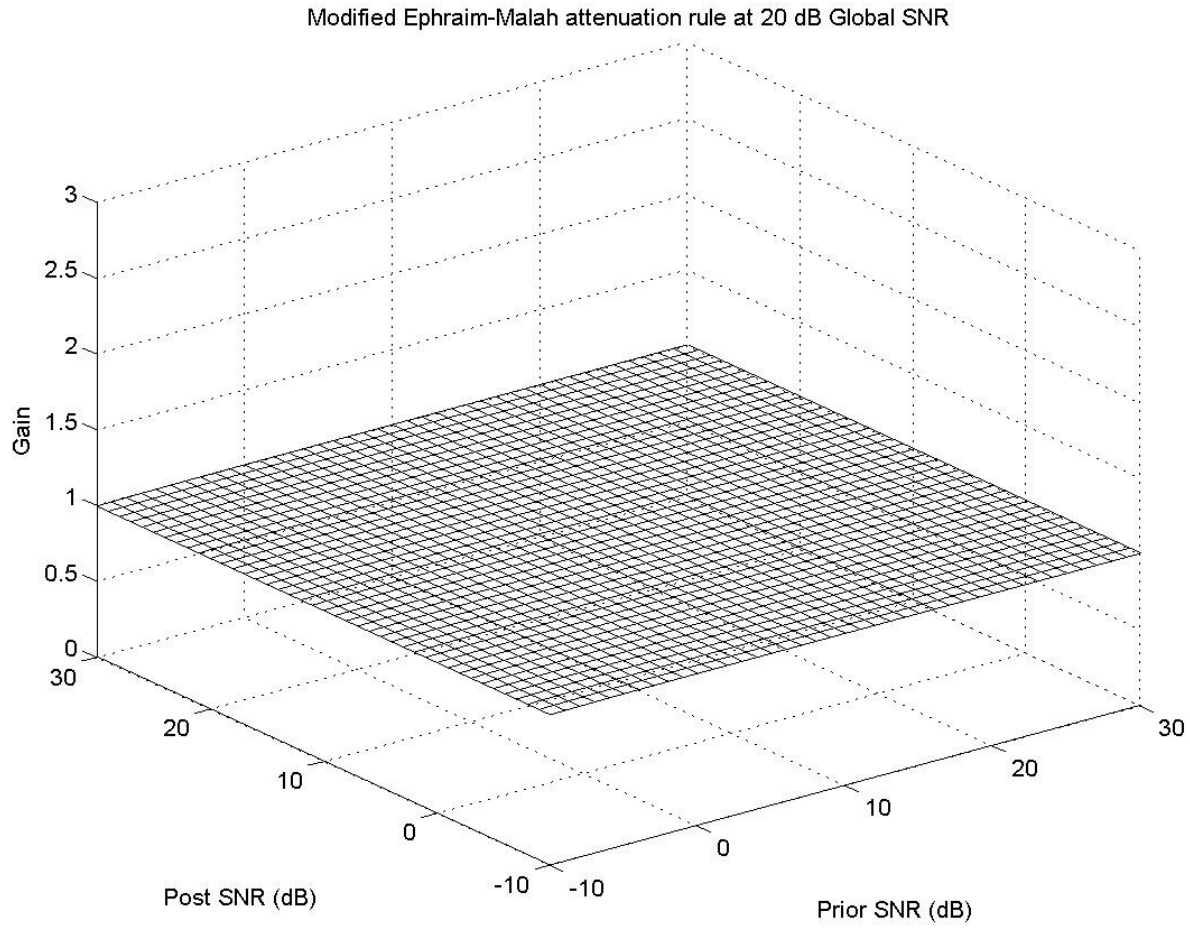


Fig. 2. Original attenuation rule computed with (4) and its version with the proposed modifications (8) and (9) at 0 dB, 5 dB, 10dB, 15dB, 20dB of Global SNR.

Table I Denoising results for Aurora-2 with different test sets and systems

	Test A		Test B		Test C		Test A-B-C Average		Overall Average
	Clean	*Multi*	*Clean*	*Multi*	*Clean*	*Multi*	*Clean*	*Multi*	
No Denoising	24.4	6.5	22.5	8.9	24.7	9.8	23.7	8.1	15.9
Wiener modified	16.0	6.1	15.6	7.9	16.7	9.5	16.0	7.5	11.8
Ephraim-Malah modified	14.7	6.0	15.8	8.0	15.2	8.9	15.2	7.4	11.3
ETSI AFE	16.1	6.4	14.7	8.2	20.2	10.1	16.4	7.9	12.1

Table II Denoising results for Aurora-2 with different test sets and systems with models trained with the Macrophone American telephone corpus

Denoising Method	Test Set A	Test Set B	Test Set C	Average
No Denoising	19.1	16.8	22.5	18.9
Wiener modified	12.1	11.9	13.6	12.3
Wiener modified	11.9	11.7	13.7	12.2
Ephraim-Malah baseline	11.6	11.9	12.7	11.9
Ephraim-Malah modified	11.0	11.4	13.1	11.6

Table III 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	43.3	30.1	17.5	**30.3**
Wiener baseline	31.9	18.7	12.2	**20.9**
Wiener modified	25.1	13.8	10.8	**16.6**
Ephraim-Malah baseline	30.3	18.7	10.3	**19.8**
Ephraim-Malah modified	24.4	12.3	9.5	**15.4**

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

Aurora 3 Test on CH1 component of train set				
A. Denoising Method	Italian C.I. 0.9	Spanish C.I. 0.6	German C.I. 1.0	V. AVERAGE
No denoising	41.1	26.2	14.2	**27.3**
Wiener baseline	29.3	18.7	10.1	**19.4**
Wiener modified	23.7	11.1	9.2	**14.7**
Ephraim-Malah baseline	28.5	14.1	9.4	**17.4**
Ephraim-Malah modified	22.5	9.4	7.9	**13.3**