



ELSEVIER

Signal Processing 75 (1999) 151–159

**SIGNAL
PROCESSING**

Improved noise suppression filter using self-adaptive estimator of probability of speech absence

Ing Yann Soon^{a,*}, Soo Ngee Koh^a, Chai Kiat Yeo^b

^a*School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 2263, Singapore*

^b*School of Applied Science, Nanyang Technological University, Nanyang Avenue, Singapore 2263, Singapore*

Received 15 January 1996; received in revised form 23 September 1998

Abstract

In this paper, two estimators of the probability of speech absence are derived using the common assumption that the Fourier coefficients of a frame of speech and noise samples are statistically independent Gaussian random variables (Ephraim and Malah, 1984; McAulay and Malpass, 1980). The estimators are obtained directly from the noisy speech itself. The first estimator is obtained by binary classification of the received spectral amplitude into speech present or speech absent state. The second estimator is obtained by deriving the conditional probability of speech absence given the received spectral amplitude. Each of the time-adaptive estimators produces an estimate of the probability of speech absence for each spectral frequency. The estimated probability will be higher during the speech period and lower during the silence period. The estimated probability can be fed directly to any filter which requires such an estimate, e.g. the Ephraim and Malah noise suppressor (Ephraim and Malah, 1984), and the modified power subtraction method (Scalart and Vieira Filho, 1996), with significant improvements for various noise types. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel werden zwei Schätzer für die Wahrscheinlichkeit des Nichtvorhandenseins eines Sprachsignals unter Verwendung der üblichen Annahme abgeleitet, daß die Fourierkoeffizienten eines Frames von Sprach- und Geräuschab-tastwerten unabhängig, gaußverteilte Zufallsvariablen sind (Ephraim und Malah, 1984; McAulay und Malpass, 1980). Die Schätzer werden direkt aus dem verrauschten Sprachsignal bestimmt. Der erste Schätzer wird durch binäre Klassifikation der Amplitude des empfangenen Spektrums in die Klassen "Sprache vorhanden" und "Sprache nicht vorhanden" erhalten. Der zweite Schätzer wird durch die Herleitung der bedingten Wahrscheinlichkeit des Nichtvorhandenseins eines Sprachsignals bei gegebenem empfangenem Amplitudenspektrum bestimmt. Jeder der zeitlich adaptiven Schätzer erzeugt einen Schätzwert für die Wahrscheinlichkeit des Nichtvorhandenseins eines Sprachsignals für jede spektrale Frequenz. Die geschätzte Wahrscheinlichkeit wird während der Sprechphasen höher sein und niedriger bei Schweigen. Die geschätzten Wahrscheinlichkeiten können direkt jedem Filter zugeführt werden, das solch eine Schätzung benötigt, beispielsweise der Ephraim und Malah Rauschunterdrücker (Ephraim und Malah, 1984) und die modifizierte Leistungsubtraktionsmethode (Scalart und Vieira Filho, 1996), was zu signifikanten Verbesserungen für verschiedenste Rauscharten führt. © 1999 Elsevier Science B.V. All rights reserved.

* Corresponding author. Tel.: + 65 790 5638; fax: + 65 792 0415; e-mail: eiyssoon@ntu.edu.sg

Résumé

Dans cet article, nous dérivons deux estimateurs de la probabilité d'absence de parole, en utilisant la supposition commune que les coefficients de Fourier des échantillons d'une trame de parole et du bruit sont des variables aléatoires gaussiennes indépendantes (Ephraim et Malah, 1984; McAulay et Malpass, 1980). Les estimateurs sont obtenus directement à partir de la parole bruitée elle-même. Le premier estimateur est obtenu par classification binaire de l'amplitude spectrale reçue en un état de présence ou d'absence de parole. Le second estimateur est obtenu en dérivant la probabilité conditionnelle d'absence de parole étant donnée l'amplitude spectrale reçue. Chacun des estimateurs adaptatifs dans le temps produit une estimation de la probabilité d'absence de parole pour chaque fréquence spectrale. La probabilité sera plus grande durant une période de parole et plus basse durant une période de silence. La probabilité estimée peut être directement entrée dans n'importe quel filtre qui nécessite une telle estimation, par exemple le supprimeur de bruit de Ephraim et Malah (1984) ou la méthode de soustraction de puissance modifiée (Scalart et Vieira Filho, 1996), produisant une amélioration significative pour différents types de bruit. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Speech enhancement; Speech processing; Noise reduction

1. Introduction

There are two forms of speech absence which can be used to improve the noise removal filter. The first form of speech absence is due to the speaker pausing in his speech resulting in significant portions of silence. The second form of speech absence is that although the speaker is talking, the speech energy is not present in all the frequency components. For some frequency components with insignificant energy, speech can be considered to be absent in those components.

Such knowledge of speech absence can be used to improve a speech enhancement filter. The first attempt at utilizing the uncertainty of speech absence was explored by McAulay and Malpass [2]. In their approach, they derived a filter based on a fixed probability of speech absence of 0.5. The Ephraim and Malah noise removal filter [1] adopted a more flexible approach in which different spectral frequency components can be assigned a different probability of speech absence which ranges from zero to one. However, the paper did not touch on how the probability of speech absence can be estimated, and for performance evaluation, the probability of speech absence was set to 0.2 empirically. Intuitively, we expect the probability of speech absence to be a function of time and frequency.

In this paper, we formulate two approaches which adaptively estimate the probabilities of speech absence in different frequency components from the noisy speech itself. In the first approach,

the noisy spectral component is hard classified or binary classified into speech presence or speech absence. In the second approach, the noisy spectral component is soft classified or statistically classified as speech absence, e.g. for a certain spectral component, the probability of speech absence is 0.6. After the classification stage, the probability of speech absence is computed using a running average of the classification results.

Both methods provide a much better estimate of speech absence which varies with time and frequency rather than a constant value. During the periods where speech is absent, the probability of speech absence will be close to one while during voiced speech the probability of speech absence will be close to zero at the pitch frequency component.

The probabilities obtained are then fed into a slightly modified form of Ephraim and Malah filter which takes into account the uncertainty of signal presence. The results show both an improvement in speech quality as well as better segmental SNR values. Similarly, the technique can be applied to other filters which require the probability of speech absence input, e.g. the modified power subtraction method using a priori SNR proposed in [3]. The results obtained also show significant improvements.

2. Ephraim and Malah noise suppression filter

This section provides a brief description of the Ephraim and Malah noise suppression filter [1],

which gives excellent results. Let the k th spectral magnitude of the speech signal, noise and noisy speech be denoted by A_k , D_k and R_k , respectively. The probability of speech absence is denoted as q_k . The k th spectral output, \hat{A}_k , of the Ephraim and Malah noise suppression filter, taking into account the uncertainty of signal presence, is given by the equation

$$\hat{A}_k = \frac{G(q_k)M(-0.5; 1; -v_k)R_k\sqrt{\pi v_k}}{2\gamma_k}, \quad (1)$$

where

$$v_k = \frac{\xi_k\gamma_k}{1 + \xi_k}, \quad (2)$$

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}, \quad (3)$$

$$\gamma_k = \frac{R_k^2}{\lambda_d(k)}, \quad (4)$$

$$\lambda_x(k) = E\{|A_k|^2\}, \quad (5)$$

$$\lambda_d(k) = E\{|D_k|^2\}, \quad (6)$$

where λ_d is the expected noise power which can be estimated using various means during the silence period [4]. $M(;;)$ in Eq. (1) is the confluent hypergeometric function defined in [5, Eq. (A.1.14)]. ξ_k and γ_k are the a priori and a posteriori signal to noise ratios, respectively. $G(q_k)$ is the additional attenuation based on the uncertainty of signal presence and is defined as follows:

$$G(q_k) = \frac{1 - q_k}{1 - q_k + q_k(1 + \xi_k)\exp(-\mu_k)}. \quad (7)$$

The a priori SNR, ξ_k , can best be estimated by the decision-directed approach [1].

$$\hat{\xi}_k(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_d(k)} + (1 - \alpha)\max[0, \gamma_k(n) - 1]. \quad (8)$$

In [1], ξ_k estimated using the above is further divided by $(1 - q_k)$. However, it is found that doing so tends to overestimate the a priori SNR especially when q_k is close to 1. Furthermore, the value of α used is 0.98 which means that the estimation is based mainly on the immediate past frame and not over a long-time period, hence the division is inap-

propriate. With this slight modification, the Ephraim and Malah noise suppressor can accept a time varying value of q_k .

3. Modified power subtraction filter

This modified power subtraction filter is a speech enhancement filter proposed by Scalart and Vieira Filho [3] using the power subtraction technique together with the a priori SNR estimated by the decision-directed approach [1] in Eq. (8). The additional attenuation for silence period can also be incorporated for a better performance. Using the same notations as in the above section, the combined filter can be described as follows:

$$\hat{A}_k = G(q_k)\sqrt{\frac{\hat{\xi}_k}{\hat{\xi}_k + 1}}R_k. \quad (9)$$

4. Hard decision estimator

The first method will be hard to classify a received noisy amplitude as one which contains speech or just noise alone. The decision will be binary, with 0 representing speech presence and 1 representing speech absence. Let the input be represented by two states, H_0 and H_1 , where H_0 : speech absence, H_1 : speech presence. Using Gaussian statistical model in [1,2], the conditional probability density function of receiving the noisy amplitude, R_k , given that speech is absent, is

$$P(R_k|H_0) = \frac{2R_k}{\lambda_d}\exp(\gamma_k). \quad (10)$$

Similarly, the conditional probability of receiving amplitude R_k given that speech is present is given by

$$P(R_k|H_1) = \frac{2R_k}{\lambda_d}\exp\left(-\frac{R_k^2 + A_k^2}{\lambda_d}\right)I_0\left(\frac{2R_kA_k}{\lambda_d}\right), \quad (11)$$

where I_0 is the zero-order modified Bessel function. However, since A_k is unknown, the approximation $\xi_k \approx A_k^2/\lambda_d$ will have to be used instead. Therefore,

$$P(R_k|H_1) \approx \frac{2R_k}{\lambda_d}\exp(-\gamma_k - \xi_k)I_0(2\sqrt{\gamma_k\xi_k}). \quad (12)$$

Using the conditional probabilities, the binary decision D_k can be obtained as follows:

If $P(R_k|H_1) > P(R_k|H_0)$ then

$$D_k = 0 \quad \{\text{speech presence}\}$$

else

$$D_k = 1 \quad \{\text{speech absence}\}.$$

Upon substitution of Eqs. (10) and (11) and simplification, the decision can also be written in the following manner:

If $\exp(-\xi_k)I_0(2\sqrt{\gamma_k\xi_k}) > 1$ then

$$D_k = 0$$

else

$$D_k = 1.$$

The probability of speech absence, q_k , can then be obtained using a running average of D_k s of previous frames. The formula used is as follows:

$$q_{k,n} = \beta D_{k,n} + (1 - \beta)q_{k,n-1}, \quad (13)$$

where n represents the current frame number and $n - 1$ represents the previous frame. The running average for the hard decision is necessary to smooth out the changes in binary values of D_k . Normally a small value for β is necessary.

5. Soft decision estimator

Unlike the Hard decision method which classifies the received amplitude into either speech presence or speech absence in a binary fashion, the soft decision method produces a value which ranges from 0 to 1 to represent the probability that the received amplitude is from a speech absence state. Using the conditional probabilities, the probability that speech is absent can be obtained from Bayes theorem:

$$P(H_0|R_k) = \frac{P(R_k|H_0)P(H_0)}{P(R_k|H_0)P(H_0) + P(R_k|H_1)P(H_1)}. \quad (14)$$

However, values of $P(H_0)$ and $P(H_1)$ are unknown a priori and an initial approximation will be to assume that they are approximately equal. Therefore, Eq. (14) can be approximated as

$$P(H_0|R_k) \approx \frac{P(R_k|H_0)}{P(R_k|H_0) + P(R_k|H_1)}. \quad (15)$$

Substituting for the conditional probabilities, Eqs. (10) and (11), the following is obtained:

$$\hat{P}(H_0|R_k) = \frac{1}{1 + \exp(-\xi_k)I_0(2\sqrt{\gamma_k\xi_k})}. \quad (16)$$

Similarly, the probability of speech absence, q_k , can then be obtained using a running average of

Table 1
SEGSNR results for white noise corrupted speech

Unprocess (dB)	EMF (dB)			MPSF (dB)		
	$q_k = 0.2$	Hard	Soft	$q_k = 0.2$	Hard	Soft
-17.3	-4.876	-2.065	-2.085	-7.037	-4.048	-3.971
-16.3	-4.22	-1.57	-1.573	-6.294	-3.367	-3.264
-15.3	-3.517	-1.063	-1.053	-5.465	-2.710	-2.626
-14.3	-2.872	-0.574	-0.563	-4.65	-2.023	-1.93
-13.3	-2.2	-0.042	0.004	-3.905	-1.373	-1.32
-12.3	-1.545	0.481	0.6	-3.149	-0.748	-0.662
-11.3	-0.944	1.003	1.071	-2.337	-0.122	-0.018
-10.3	-0.242	1.549	1.68	-1.603	0.454	0.564
-9.3	0.428	2.062	2.19	-0.842	1.079	1.231
-8.3	1.143	2.557	2.755	-0.123	1.765	1.941
-7.3	1.81	3.092	3.342	0.664	2.418	2.619

Table 2
SEGSNR results for fan noise corrupted speech

Unprocess (dB)	EMF (dB)			MPSF (dB)		
	$q_k = 0.2$	Hard	Soft	$q_k = 0.2$	Hard	Soft
– 15.8	– 1.144	2.373	2.104	– 2.597	1.337	0.456
– 14.8	– 0.351	2.991	2.758	– 1.763	2.076	1.2
– 13.8	0.442	3.619	3.403	– 0.930	2.788	1.951
– 12.8	1.239	4.23	4.059	– 0.097	3.479	2.709
– 11.8	2.039	4.83	4.741	0.736	4.176	3.472
– 10.8	2.844	5.439	5.425	1.567	4.841	4.239
– 9.8	3.65	6.041	6.127	2.396	5.508	5.008
– 8.8	4.456	6.655	6.863	3.222	6.187	5.774
– 7.8	5.261	7.285	7.591	4.045	6.849	6.536
– 6.8	6.063	7.934	8.323	4.865	7.574	7.295
– 5.8	6.861	8.624	9.044	5.684	8.285	8.049

Table 3
SEGSNR results for F16 noise corrupted speech

Unprocess (dB)	EMF (dB)			MPSF (dB)		
	$q_k = 0.2$	Hard	Soft	$q_k = 0.2$	Hard	Soft
– 17.16	– 4.612	– 3.383	– 3.018	– 7.839	– 5.476	– 5.036
– 16.16	– 4.034	– 2.884	– 2.524	– 7.067	– 4.82	– 4.395
– 15.16	– 3.451	– 2.373	– 2.014	– 6.298	– 4.164	– 3.756
– 14.16	– 2.868	– 1.861	– 1.509	– 5.531	– 3.505	– 3.115
– 13.16	– 2.284	– 1.35	– 0.994	– 4.765	– 2.852	– 2.468
– 12.16	– 1.686	– 0.82	– 0.464	– 4.0	– 2.203	– 1.811
– 11.16	– 1.077	– 0.273	0.0857	– 3.234	– 1.545	– 1.148
– 10.16	– 0.467	0.275	0.653	– 2.468	– 0.869	– 0.481
– 9.16	0.152	0.831	1.239	– 1.7	– 0.203	0.191
– 8.16	0.782	1.398	1.826	– 0.939	0.467	0.87
– 7.16	1.42	1.971	2.429	– 0.163	1.138	1.557

$\hat{P}(H_0|R_k)$ of previous frames. The formula used is as follows:

$$q_{k,n} = \beta \hat{P}(H_0|R_k)_n + (1 - \beta)q_{k,n-1}, \quad (17)$$

where n represents the current frame number and $n - 1$ represents the previous frame number.

6. Results and discussions

A total of 10 different utterances, taken from the TIMIT database, are used in our evaluation. Half

of the utterances were from male speakers while the rest are from female speakers. The speech data used are sampled at 8 kHz and quantized linearly using 16 bits. As for the additive noise, three different noise types were used, namely Gaussian white noise, recorded fan noise as well as the F16 (fighter jet) noise from the NOISEX database.

The noisy speech data were divided into frames, each of which consists of 256 samples with an overlap of 192 samples with the neighbouring frame. Hanning windowing is performed on each frame before it is enhanced individually. The final

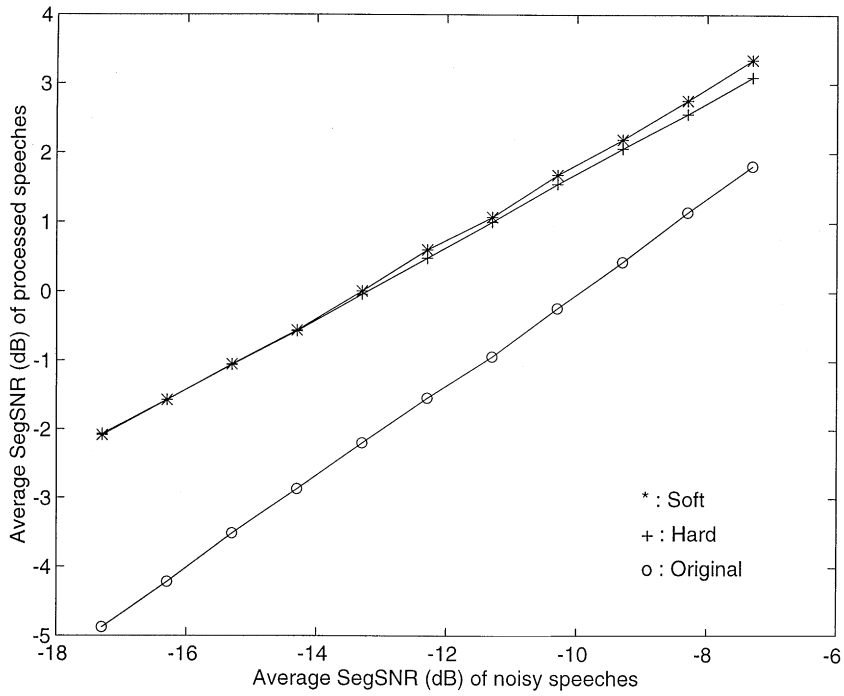


Fig. 1. Results for white noise corrupted speech processed by EMF.

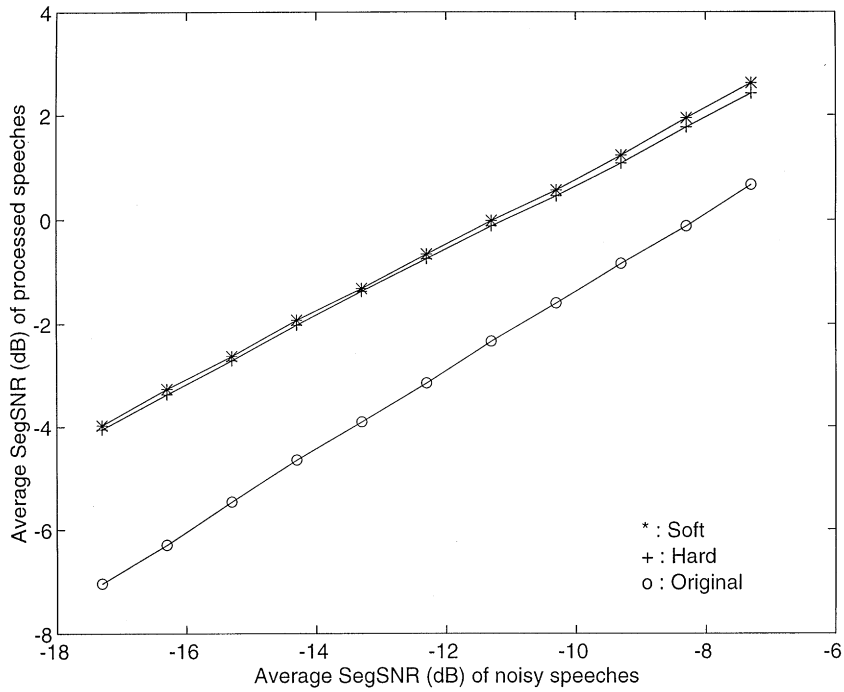


Fig. 2. Results for white noise corrupted speech processed by MPFSF.

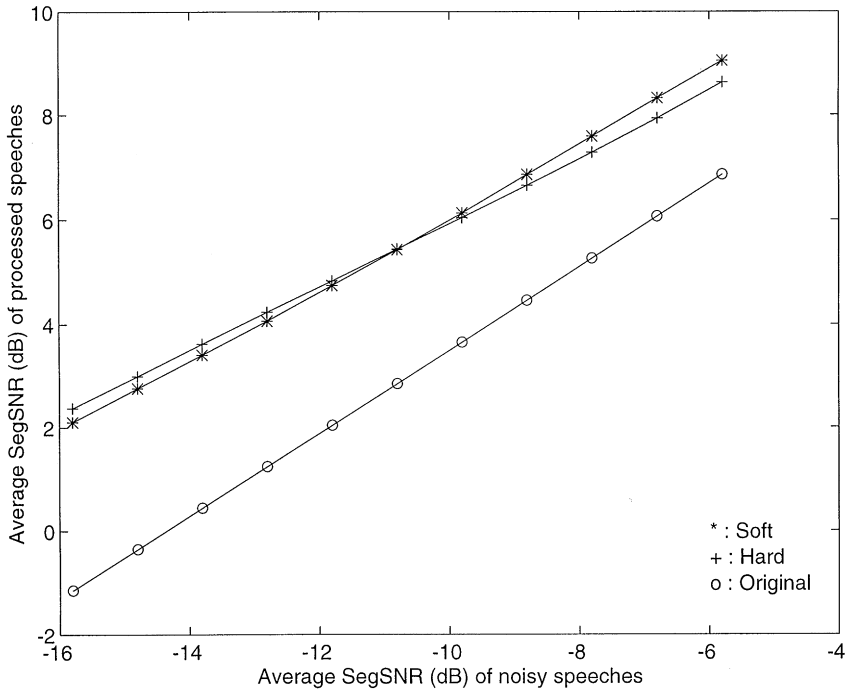


Fig. 3. Results for fan noise corrupted speech processed by EMF.

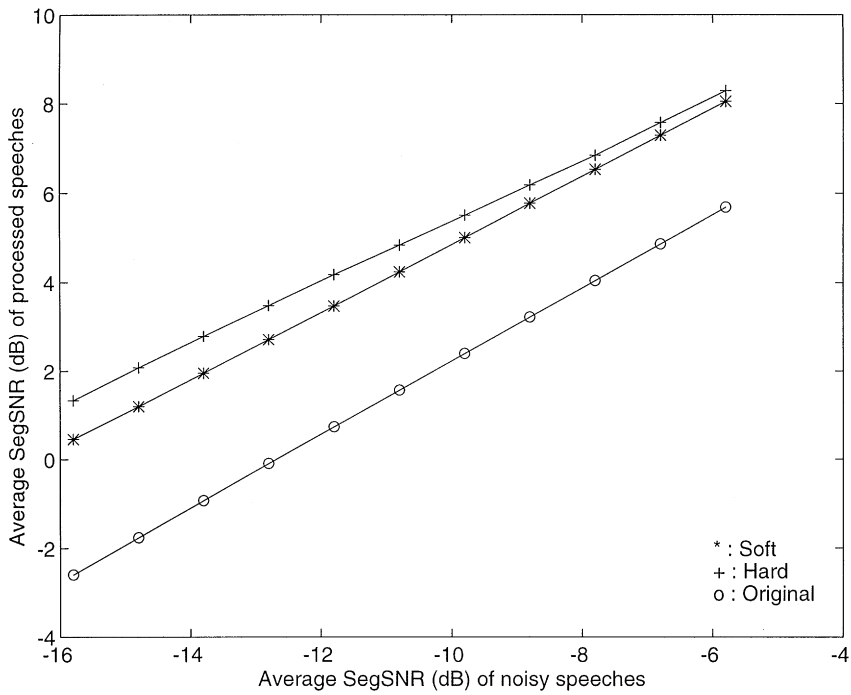


Fig. 4. Results for fan noise corrupted speech processed by MPSF.

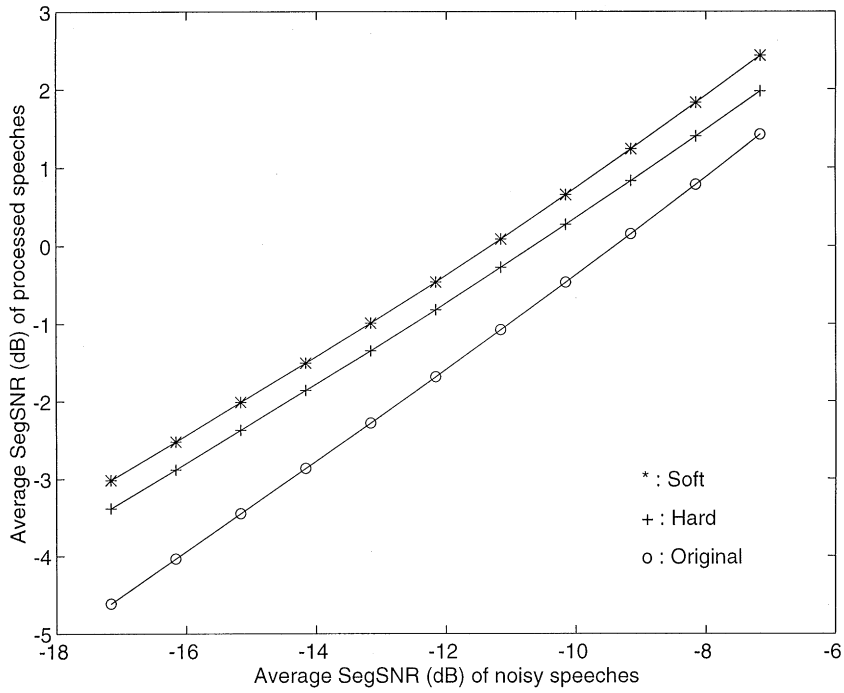


Fig. 5. Results for F16 noise corrupted speech processed by EMF.

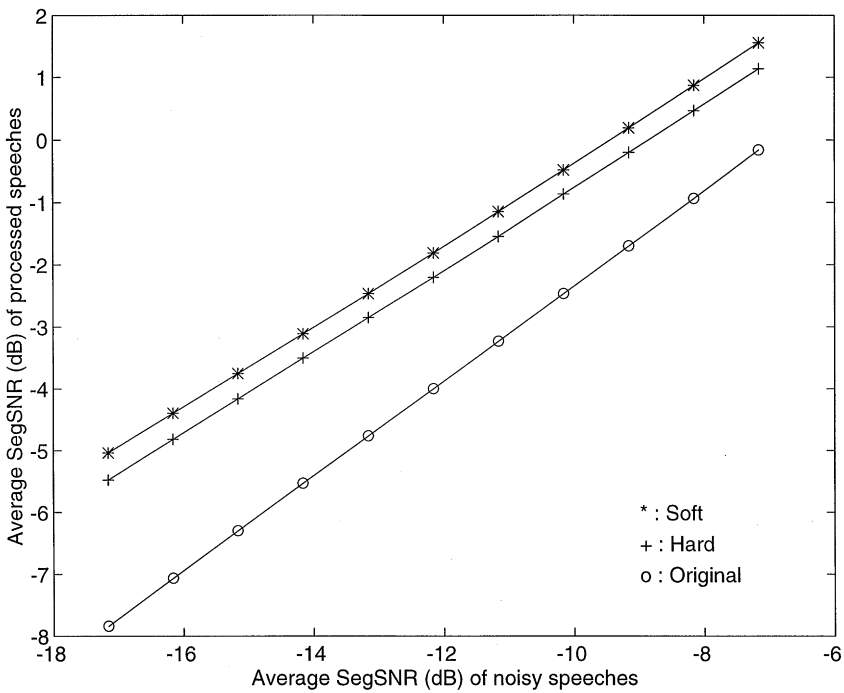


Fig. 6. Results for F16 noise corrupted speech processed by MPSF.

enhanced speech is reconstructed from the enhanced frames using the weighted overlap and add technique [6]. The values of α in Eq. (8) and β in Eq. (13) are set to 0.98 and 0.1, respectively.

The same amount of noise was added to all the 10 utterances and the average segmental SNR is computed before and after filtering. The results for various noise type are tabulated in Tables 1–3. Tables 1–3 contain the results for white noise, fan noise and F16 noise, respectively. The first column in the table contains the average segmental SNR of the unprocessed speech while the next three columns contain the average segmental SNR values of the filtered speeches using various implementations of the Ephraim and Malah noise suppressor (EMF). The first implementation is to use a static value of q_k as suggested by Ephraim and Malah. The static value of q_k was empirically chosen to be 0.2 according to [1]. The second and third implementations involve the use of the Hard and Soft decision schemes to obtain q_k as presented in Sections 4 and 5. The last three columns are the results for the modified power subtraction (MPSF) case.

The results are also illustrated graphically for easier viewing in Figs. 1–6. It can be clearly seen that significant improvement in segmental SNR can be obtained using either the Hard or Soft decision schemes as compared to the original implementations. However, the choice between the soft and hard decision estimator is not clear cut, as the result seems to be noise and filter dependent. The soft decision gives better results for white noise and F16 noise. However, for fan noise using MPSF, the hard decision is always better.

From listening tests, both estimators result in significantly lower residual noise as compared to

the original implementation with fixed probability of speech absence. The soft decision estimator seems to sound slightly better than the hard decision estimator in most cases.

7. Conclusions

This paper proposes two methods of estimating the probability of speech absence adaptively from the noisy speech itself. It shows that by using these two estimators of speech absence, the performance of the Ephraim and Malah noise filter and the power subtraction filter can be significantly improved. The technique should also be applicable in other filters incorporating the probability of speech absence.

References

- [1] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-32 (December 1984) 1109–1121.
- [2] R.J. McAulay, M.L. Malpass, Speech enhancement using a soft-decision noise suppression filter, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-28 (April 1980) 137–145.
- [3] P. Scalart, J. Vieira Filho, Speech enhancement based on a priori signal to noise estimation, in: *Proc. ICASSP*, Vol. 2, 1996, pp. 629–632.
- [4] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-27 (April 1979) 113–120.
- [5] I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, 1980.
- [6] R.E. Crochiere, A weighted overlap-add method of short-time Fourier analysis/synthesis, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-28 (February 1980) 99–102.