Speech Enhancement Using Minimum Mean-Square Error Amplitude Estimators Under Normal and Generalized Gamma Distribution

[1,2]Chabane Boubakir and [2]Daoud Berkani
[1]Lamel Laboratory, Department of Electronic, Jijel University,
BP 98, Ouled Aissa, JIJEL, Algeria
[2]Departments of Signal and Communications, National Polytechnic School,
10 Avenue Hassen Badi, BP 182, El-Harrach, 16200, Algeria

Abstract: Problem statement: In this study, DFT-based speech enhancement via Minimum Mean-Square Error (MMSE) amplitude estimators was considered. **Approach:** Several variants of the basic approach (MMSE-STSA) have been proposed over the years to address certain shortcomings, chiefly the quality of the remnant noise and its trade-off with speech distortion. In this study, we presented a comparative study between the MMLSA and the estimators based on the Gamma model, followed by an implementation in Matlab of these algorithms and an objective evaluation using a corpus of speech. **Results:** We obtained the best values of various parameters used by different estimators. **Conclusion:** Objective evaluation confirm superiority in noise suppression and quality of the enhanced speech by the estimators derived under the generalized Gamma distribution than the estimators derived under the normal distribution, in stationary environments.

Key words: Speech enhancement, DFT-based speech enhancement, MMSE estimators, Gamma distribution

INTRODUCTION

The interest in the field of speech enhancement emerges from the increased usage of digital speech processing applications like mobile telephony, digital hearing aids and human-machine communication systems in our daily life. The trend to make these applications mobile increases the variety of potential sources for quality degradation. Speech enhancement methods can be used to increase the quality of these speech processing devices and make them more robust under noisy conditions. The large group of speech enhancement methods meant to improve certain quality aspects of these devices. In this study we will focus on single-microphone additive noise reduction and aim at methods that study in the Discrete Fourier Transform (DFT) domain.

The traditional hypothesis for speech enhancement in the DFT domain is that the distribution of the complex speech DFT coefficients is Gaussian (Ephraim and Malah, 1984; 1985). Therefore, the spectral amplitude distribution is modeled by a Rayleigh distribution. Actually, super-Gaussian models of the DFT coefficients are used, because they lead to estimators with improved performance than those based on a Gaussian model. Martin (2005) derived complex-DFT estimators under Laplacian and Gamma speech assumptions. Lotter and Vary (2005) proposed a MAP amplitude estimator for a generalized Gamma amplitude distribution.

MMSE estimators of the complex DFT coefficients, assuming a two-sided generalized Gamma distribution, have been derived in (Jensen *et al*., 2006). MMSE estimators for the amplitudes, assuming a one-sided generalized Gamma distribution, are treated in (Andrianakis and White, 2006) and (Hendriks *et al*., 2006). For all these estimators, the decision-directed method is commonly used (Ephraim and Malah, 1984).

In this study, we present a comparative study between the MMLSA, which is the most efficient variant of the estimators based on the Gaussian model and the estimators based on the Gamma model. This study is followed by an implementation in Matlab of these algorithms and an objective evaluation using a corpus of speech.

Corresponding Author: Chabane Boubakir, Lamel Laboratory, Department of Electronic, Jijel University, BP 98, Ouled Aissa, JIJEL, Algeria

MATERIALS AND METHODS

MMSE spectral estimation:
Modeling noise DFT magnitudes and assumptions:
Assume that we observe a noisy speech signal y(t) that is a sum of a speech and noise signal x(t) and d(t), which are uncorrelated. Their representation in the Short Time Fourier Transform (STFT) domain is given by:

$$Y(k,l) = X(k,l) + D(k,l) \qquad (1)$$

where, Y(k, 1) and D(k, 1) are the samples of the noisy speech, the clean speech and the noise signal's STFT correspondingly. The index k corresponds to the frequency bins and the index l to the time frames of the STFT. Since DFT coefficients from different time frames and frequency indices are assumed to be independent, the indices k and l will be omitted for simplicity. We can write $X = Ae^{j\Phi}$ and $Y = Ae^{j\Phi}$, where random variables A and R represent the clean and noisy amplitude and Φ and Θ the corresponding phases values.

In this study we focus on MMSE estimation of the clean amplitude A. The MMSE estimate of A is the expectation of the clean amplitude conditional on the noisy amplitude r(E{A/r}). With Bayes formula we can express the MMSE estimate $\hat{A}$ as:

$$\hat{A} = E\{A/r\} = \frac{\int_0^\infty a f_{R/A}(r/a) f_A(a) da}{\int_0^\infty f_{R/A}(r/a) f_A(a) da} \qquad (2)$$

The estimation of the clean amplitude A requires some assumptions about the distribution of the speech and the noise. The speech has usually been assumed Gaussian, e.g., (Ephraim and Malah, 1984; 1985), but in recent times estimators based on super-Gaussian speech assumptions such as Laplacian or Gamma distributions have been derived (Lotter and Vary, 2004). A similar development has been seen for the noise assumptions; most commonly the noise is assumed Gaussian, but estimators exist which suppose the noise to obey a super-Gaussian distribution (Lotter and Vary, 2004).

With the zero-mean Gaussian distribution assumption of the noise DFT coefficients, $f_{R/A}(r/a)$ can be written as (McAulay and Malpass, 1980):

$$f_{R/A}(r/a) = \frac{2r}{\sigma_D^2} \exp\left(-\frac{r^2+a^2}{\sigma_D^2}\right) I_0\left(\frac{2ar}{\sigma_D^2}\right) \qquad (3)$$

Where:

I_o = The 0th order modified Bessel function of the first kind

$\sigma_D^2 = E\left\{|D|^2\right\}$ = The noise spectral variance

Gaussian based short-time spectral amplitude estimator: In this case, the DFT coefficients of both the speech and the noise are assumed to be an independent Gaussian random variables. Moreover, the speech signal might not be present at all times and at all frequencies. We therefore consider a two following hypotheses H_0^k and H_1^k :

H_0^k = Speech absent in kth DFT bin, $Y_k = D_k$
H_1^k = Speech present in kth DFT bin, $Y_k = X_k + D_k$

Hence, the probability density function can be given as:

$$P\left(Y_k / H_0^k\right) = \frac{1}{\pi \lambda_d(k)} \exp\left(\frac{-|Y_k|^2}{\lambda_d(k)}\right) \qquad (4)$$

$$P\left(Y_k / H_1^k\right) = \frac{1}{\pi(\lambda_x(k)+\lambda_d(k))} \exp\left(\frac{-|Y_k|^2}{\lambda_x(k)+\lambda_d(k)}\right) \qquad (5)$$

where, $\lambda_x(k) = E\left\{|X_k|^2 / H_1^k\right\}$ and $\lambda_d(k) = E\left\{|D_k|^2\right\}$ are the variance of the spectral component of speech and noise.

Let C_k be some function of the short-time spectral amplitude A_k of the clean speech in the kth bin (e.g., A_k , $\log A_k$, A_k^2). The MMSE estimator $\hat{C}_k$ of C_k is given by (McAulay and Malpass, 1980):

$$C_k = E\left\{C_k / Y_k, H_1^k\right\} P\left(H_1^k / Y_k\right) + E\left\{C_k / Y_k, H_0^k\right\} P\left(H_0^k / Y_k\right) \qquad (6)$$

where, $E\{./.\}$ and $P(./.)$ denote conditional expectations and conditional probabilities, respectively. Since $E\left\{C_k / Y_k, H_0^k\right\} = 0$, we have:

$$C_k = E\left\{C_k / Y_k, H_1^k\right\} P\left(H_1^k / Y_k\right) \qquad (7)$$

Thus, $G_{MM}(k) = P\left(H_1^k / Y_k\right)$ is the Multiplicatively-Modification of the optimal estimator under the speech presence hypothesis $\left(G_{LSA}(k) = E\left\{C_k / Y_k, H_1^k\right\}\right)$.

Based on the results reported in (Ephraim and Malah, 1985; Malah *et al*., 1999), the Multiplicatively-Modified Log-Spectral Amplitude (MM-LSA) estimator (corresponding to $C_k = \log A_k$) outperformed the traditional MMSE-STSA estimator (Ephraim and Malah, 1984) with and without incorporating speech presence uncertainty indicated as MMSE-SPU and MMSE respectively ($C_k = A_k$).

The MM-LSA estimator is (Malah *et al*., 1999):

$$\hat{A}_{MMLSA} = G_{MM}(k)G_{LSA}(k)R_k \tag{8}$$

Under the Gaussian assumptions on the speech and noise, the gain function $G_{LSA}(k)$ is derived in (Ephraim and Malah, 1985) to be:

$$G_{LSA}(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{9}$$

Where:

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$$

$$\gamma_k = \frac{R_k^2}{E\{|D_k|^2\}}$$

$$\xi_k = \frac{\eta_k}{1-q_k}$$

$$\eta_k = \frac{E\{|X_k|^2\}}{E\{|D_k|^2\}}$$

with η_k is called the a priori SNR, γ_k is the a posteriori SNR and q_k is the a prior probability of speech absence in the k-th bin.

The gain modification $G_{MM}(k)$ is the soft-decision modification of the optimal estimator under the signal presence hypothesis and is given by (Ephraim and Malah, 1984; Malah *et al*., 1999):

$$G_{MM}(k) = \frac{\Lambda(k)}{1+\Lambda(k)} \tag{10}$$

where the likelihood ratio $\Lambda(k)$ is defined as:

$$\Lambda(k) = \frac{P(H_1^k)}{P(H_0^k)} \cdot \frac{P(Y_k/H_1^k)}{P(Y_k/H_0^k)} = \frac{(1-q_k)}{q_k} \cdot \frac{P(Y_k/H_1^k)}{P(Y_k/H_0^k)}$$

and, q_k denotes the a priori probability of speech absence in the kth bin. By using 4 and 5, we get:

$$\Lambda(k) = \frac{(1-q_k)}{q_k} \frac{\exp(v_k)}{1+\xi_k}\bigg|_{\xi_k = \eta_k/(1-q_k)} \tag{11}$$

Gamma based short-time spectral amplitude estimator.

In the Gamma based MMSE estimators of the speech DFT magnitudes; we assume that the speech DFT magnitudes are distributed according to a one-sided generalized Gamma prior density of the form:

$$f_A(a) = \frac{\gamma\beta^v}{\Gamma(v)} a^{\gamma v-1} \exp(-\beta a^\gamma), \quad a \geq 0 \tag{12}$$

where $\Gamma(.)$ is the Gamma function and the random variable A represents the DFT magnitudes, with the constraints on the parameters $\beta>0$, $\gamma>0$, $v>0$.

The Gamma based MMSE amplitude estimators for the cases $\gamma = 1$ and $\gamma = 2$ have been derived in (Andrianakis and White, 2006; Hendriks *et al*., 2006; Erkelens *et al*., 2007). We will use the case $\gamma = 2$, as the related estimator can be derived without any approximations and the maximum achievable performance for both cases is about the same.

Inserting Eq. 12 with $\gamma = 2$ and Eq. 3 into 2 gives:

$$\hat{A}^{(2)} = \frac{\int_0^\infty a^{2v} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a^2\right) I_0\left(\frac{2ar}{\sigma_D^2}\right) da}{\int_0^\infty a^{2v-1} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a^2\right) I_0\left(\frac{2ar}{\sigma_D^2}\right) da} \tag{13}$$

where, the superscript (2) indicates that the parameter $\gamma = 2$. Using (Gradshteyn and Ryzhik, 2000), the integrals can be solved for $v>0$. After inserting the relation between β and the second moment $E\{A^2\}$, which for this case is $\beta = v/\sigma_x^2$, with $\sigma_x^2 = E\{|X_k|^2\}$, the estimator is (Hendriks *et al*., 2006):

$$\hat{A}_k^{(2)} = \frac{\Gamma(v+0.5)}{\Gamma(v)} \frac{\sqrt{Q_k}}{\gamma_k} \frac{M_{(-v,0)}(Q_k)}{M_{(-v+0.5,0)}(Q_k)} r \tag{14}$$

where, $Q_k = \gamma_k \eta_k / (v + \eta_k)$, $M_{v,\mu}$ is recognized as the Whittaker function, or in terms of confluent hypergeometric function $_1F_1(a;b;x)$ (Gradshteyn and Ryzhik, 2000):

$$\hat{A}_k^{(2)} = \frac{\Gamma(v+0.5)}{\Gamma(v)} \frac{\sqrt{Q_k}}{\gamma_k} \frac{_1F_1(v+0.5;1;Q_k)}{_1F_1(v;1;Q_k)} r \tag{15}$$

The special case $\nu = 1$ is the traditional MMSE-STSA estimator derived in (Ephraim and Malah, 1984).

The decision-directed estimator of the a priori SNR:
In order to evaluate the above gain functions, we must first estimate the noise power spectrum $\lambda_d(k) = E\{|D_k|^2\}$. This is often done during periods of speech absence as determined by a Voice Activity Detector (VAD), by using a noise-estimation algorithm like the minimum statistics approach (Martin, 1994; 2001), or by using a real noise in comparative studies.

The a posteriori SNR estimator γ_k is the ratio of the squared input amplitude R_k^2 and the estimated noise spectrum.

In (Ephraim and Malah, 1984; 1985; Cape, 1994), a decision-directed approach for the a priori SNR estimation is proposed:

$$\hat{\eta}_k(1) = \max\left[\alpha\frac{\hat{A}^2(k,1-1)}{\lambda_d(k,1-1)}+(1-\alpha)\left[\gamma(k,1)-1\right],\eta_{min}\right] \quad (16)$$

where the smoothing factor $0\leq\alpha\leq1$, a value of $\alpha = 0.98$ was used in the implementation and the lower limit η_{min} recommended by (Cape, 1994), is the same to the use of the spectral floor in the basic spectral subtraction method (Berouti *et al.*, 1979). A lower limit of at least-15 dB is recommended.

Implementation and performance evaluation: For the experiment, the Noizeus database (Hu and Loizou, 2007) was used which consists of 30 IRS-filtered speech signals sampled at 8 kHz, contaminated by various additive noise sources. The frame size is 256 samples, with an overlap of 50%. The data window used was a Hanning window. The enhanced signal was combined using the overlap and add approach. The a priori probability of speech absence, q_k, was set to $q_k = 0.3$ in (7). The noise variance was estimated from 0.64 seconds of noise only, preceding speech activity. Matlab implementations available from, (Borrowes, 2003) have been used to evaluate the confluent hypergeometric functions.

To measure quality of the enhanced signal, we have used the segmental SNR, the Log-Likelihood Ratio measure (LLR) (Hansen and Pellom, 1998) and the Perceptual Evaluation of Speech Quality (PESQ) (Rix *et al.*, 2001). All the measures show high correlation with subjective quality.

The LLR measure for each 20-ms speech frame is given by:

$$d_{LLR}\left(\vec{a}_d, \vec{a}_j\right) = \log\left(\frac{\vec{a}_d R_j \vec{a}_d^T}{\vec{a}_j R_j \vec{a}_j^T}\right) \quad (17)$$

Where:

$\vec{a}_\phi$ and $\vec{a}_d$ = The Linear Prediction (LP) coefficient vectors for the clean and processed speech frame respectively

R_ϕ = The autocorrelation of the clean speech frame

LLR = A spectral distance measure which mainly models the mismatch between the formants of the original and enhanced signals

The mean LLR value was obtained by averaging the individual frame LLR values across the sentence. The highest 5% of the LLR measures values were discarded, as suggested in (Hansen and Pellom, 1998), to exclude unrealistically high spectral distance values. The lower LLR measures for an enhanced speech, the better are its perceived quality.

Since the correlation of SNR with subjective quality is so poor. Instead, we choose the frame-based segmental SNR by averaging frame level SNR estimates and is defined by (Hansen and Pellom, 1998):

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\displaystyle\sum_{t=Nm}^{Nm+N-1}s_\phi^2(n)}{\displaystyle\sum_{t=Nm}^{Nm+N-1}\left[s_d(n)-s_\phi(n)\right]^2} \quad (18)$$

where, M denotes the number of frames. The lower and upper thresholds are selected to be -10 dB and +35 dB, respectively.

The perceptual evaluation of speech quality (Rix *et al.*, 2001), predicts the subjective quality of speech signals with high correlation between subjective and objective results and expresses the quality in a score from 1.0 (worst) up to 4.5 (best).

RESULTS AND DISCUSSION

We evaluate the two estimators (MM-LSA and the Gamma based estimator). For a proper choice of ν, we evaluated the estimator for a wide range of values between 0.01 and 2.5. Figure 1 and 2 shows plots of SNR segmental and PESQ versus for $\gamma = 2$, at 0 and 5 dB SNR, in the case of white noise and babble noise, respectively.

We see the similarity between the PESQ plots and the SNR seg plots. Furthermore, the better performance is reached with lower ν-values and the Gamma based estimator scores very well for $\nu\approx0.1$.
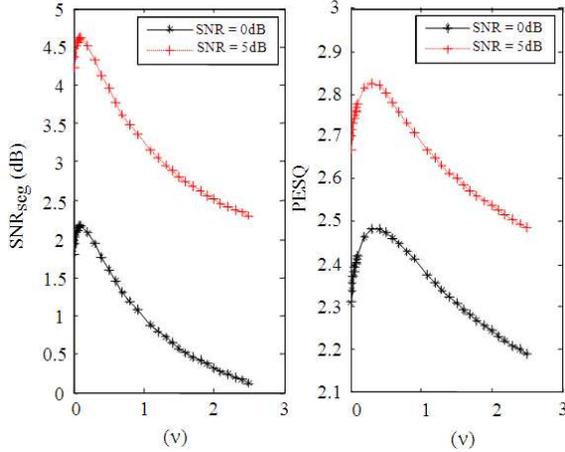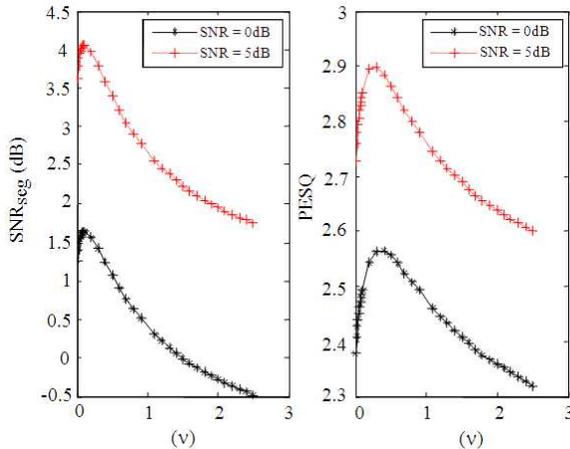
Fig. 1: SNR segmental and PESQ performance

Fig. 2: SNR segmental and PESQ performance versus ν for Babble noise

Table 1: Objective quality scores for various algorithms under white, babble and car noise, SNR = 5 dB

Noise	Method	5 dB SNR		
		LLR	SNR_{seg}	PESQ
White	Noisy	1.545	-2.327	1.799
	MMLSA	0.945	4.204	2.706
	GAMMA	0.891	4.609	2.775
Babble	Noisy	0.715	-1.783	2.006
	MMLSA	0.478	3.661	2.791
	GAMMA	0.437	4.059	2.851
Car	Noisy	0.795	-2.173	1.891
	MMLSA	0.511	3.589	2.703
	GAMMA	0.469	3.977	2.763

Table 2: Objective quality scores for various algorithms under white, babble and car noise, SNR = 0 dB

Noise	Method	0 dB SNR		
		LLR	SNR_{seg}	PESQ
White	Noisy	1.802	-5.081	1.539
	MMLSA	1.142	1.822	2.351
	GAMMA	1.087	2.169	2.421
Babble	Noisy	0.895	-4.632	1.705
	MMLSA	0.617	1.288	2.441
	GAMMA	0.577	1.639	2.494
Car	Noisy	1.014	-4.959	1.634
	MMLSA	0.652	1.273	2.365
	GAMMA	0.612	1.609	2.426

CONCLUSION

This study considered DFT based techniques for single channel speech enhancement. We show an increase in the quality of the enhanced speech with different noise types. Results, in terms of objective measures and listening test, indicated that the Gamma based estimator yielded better performance than the MM-LSA estimator based on a Gaussian model.

In the future, we plan to evaluate its possible application in preprocessing for new communication systems and hearing aid system.

REFERENCES

Andrianakis, I. and P.R. White, 2006. MMSE speech spectral amplitude estimators with chi and gamma speech priors. Proc. ICASSP, 3: 1068-1071.

Berouti, M., R. Schwartz and J. Makhoul, 1979. Enhancement of speech corrupted by acoustic noise. Proceeding of the ICASSP, Apr. 1979, Bolt Beranek and Newman Inc., Cambridge MA., pp: 208-211.

Borrowes, B., 2003. Matlab routines for computation of special functions. http://ceta.mit.edu/ceta/comp_spec_func/

Cape, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Proc., 2: 345-349. DOI: 10.1109/89.279283

The quality of speech enhanced by the Gamma estimator was compared against the quality of speech produced by the other MMSE STSA estimators.
Table 1 and 2 summarize the objective results for noisy speech, for enhanced speech with MM-LSA estimator and for enhanced speech with Gamma based estimator.

From the results in Table 1 and 2, it can be seen that the Gamma based estimator had higher preference scores compared to the MM-LSA estimator for all noise at 0 and 5dB SNR. Further, the enhanced speech from the Gamma estimator sounds less musical than that obtained from the other estimator. This was due to the fact that the Gamma priors fit better to measured speech DFT distributions than the Gaussian priors.

Ephraim, Y. and D. Malah, 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acous. Speech Signal Proc., 32: 1109-1121.

Ephraim, Y. and D. Malah, 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Proc., 33: 443-445.

Erkelens, J.S., J. Jensen and R. Heusdens, 2007. Improved speech spectral variance estimation under the generalized gamma distribution. Proceeding of the SPS-DARTS, 3rd Annual IEEE Benelux/DSP Valley Signal Processing Symposium, (VSPS'07), IEEE, USA., pp: 43-46.

Gradshteyn, I. and I. Ryzhik, 2000. Table of Integrals, Series and Products. 6th Edn., ISBN: 0-12-294757-6, pp: 1163.

Hansen, J. and B. Pellom, 1998. An effective quality evaluation protocol for speech enhancement algorithm. Proceeding of the International Conference on Spoken Language Processing, Dec. 1998, Sydney, Australia, pp: 2819-2822. http://home.iitk.ac.in/~rhegde/ee627_2009/bhatti3.pdf

Hendriks, R.C., J.S. Erkelens, J. Jensen and R. Heusdens, 2006. Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution. Proceeding of the IWAENC, Sept. 12-14, Paris, pp: 1-4. http://www.iwaenc.org/proceedings/2006/pdf/A35.pdf

Hu, Y. and P.C. Loizou, 2007. Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun., 49: 588-601. http://www.utdallas.edu/~loizou/speech/noiseus/

Jensen, J., R.C. Hendriks, J.S. Erkelens and R. Heusdens, 2006. MMSE estimation of complex-valued discrete Fourier coefficients with generalized gamma priors. Proceeding of the Interspeech, Sept. 2006, pp: 257-260.

Lotter, T. and P. Vary, 2004. Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-Gaussian speech modeling. Proceeding of the 12th European Signal Processing Conference, (ESPC'04), Vienna, Austria, pp: 1457-1460.

Lotter, T. and P. Vary, 2005. Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Applied Signal Proc., 2005: 1110-1126. DOI: 10.1155/ASP.2005.1110

Malah, D., R. Cox and A. Accardi, 1999. Tracking speech presence uncertainty to improve speech enhancement in non stationary noise environments. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 15-19, IEEE Xplore Press, Phoenix, Ariz, pp: 789-792. DOI: 10.1109/ICASSP.1999.759789

Martin, R., 1994. Spectral subtraction based on minimum statistics. Proceeding of the European Signal Processing Conference (EUSIPCO'94), pp: 1182-1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process., 9: 504 -512.

Martin, R., 2005. Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors. IEEE Trans. Speech Audio Process., 13: 845-856. DOI: 10.1109/TSA.2005.851927

McAulay, R.J. and M.L. Malpass, 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Proc., 28: 137-145.

Rix, A., J. Beerends, M. Hollier and A. Hekstra, 2001. Perceptual Evaluation Of Speech Quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proc. IEEE ICASP, 2: 749-752. DOI: 10.1109/ICASSP.2001.941023