Journal of Universal Computer Science, vol. 24, no. 9 (2018), 1271-1281
submitted: 30/12/17, accepted: 30/6/18, appeared: 28/9/18 © J.UCS

Machine Learning Optimization of Parameters for Noise Estimation

Yuyong Jeon
(Inha University, Incheon, Korea
nicejyy@gmail.com)

Ilkyeun Ra
(University of Colorado Denver, Colorado, USA
Ilkyeun.Ra@ucdenver.edu)

Youngjin Park
(Korea Electrotechnology Research Institute, Ansan, Korea
yjpark@keri.re.kr)

Sangmin Lee
(Inha University, Incheon, Korea
sanglee@inha.ac.kr)

Abstract: In this paper, a fast and effective method of parameter optimization for noise estimation is proposed for various types of noise. The proposed method is based on gradient descent, which is one of the optimization methods used in machine learning. The learning rate of gradient descent was set to a negative value for optimizing parameters for a speech quality improvement problem. The speech quality was evaluated using a suite of measures. After parameter optimization by gradient descent, the values were re-checked using a wider range to prevent convergence to a local minimum. To optimize the problem's five parameters, the overall number of operations using the proposed method was 99.99958% smaller than that using the conventional method. The extracted optimal values increased the speech quality by 1.1307%, 3.097%, 3.742%, and 3.861% on average for signal-to-noise ratios of 0, 5, 10, and 15 dB, respectively.

Keywords: Noise Estimation, Optimization, Machine Learning, Gradient Descent
Categories: G.1.6, I.5.4

1 Introduction

For speech quality improvement, noise estimation algorithms are required for determining the noise power in noisy speech data as well as noise reduction algorithms based on the estimated noise power. Spectral subtraction, proposed by Boll in [Boll, 79], is one of the most popular noise reduction strategies. The algorithm substantially reduces the noise component in noisy speech. The most important component of spectral subtraction is the estimation of noise power, or noise component, of data. To estimate the noise power spectrum, Martin [Martin, 01] proposed a noise power spectral density estimation algorithm based on the minimum statistic, which determines the minimum of the smoothed power spectrum of noisy

speech data in a sub-window. Then, the determined minimum is used for calculating the signal-to-noise ratio (SNR), which in turn is used as a criterion for deciding the presence of speech in data. This algorithm has a drawback; the speech component in the data is attenuated following a sudden increase in the noise energy level, owing to the very slow update rate of noise estimation. To alleviate this drawback, Cohen [Cohen, 02] proposed the minima controlled recursive averaging (MCRA) algorithm, which estimates the noise spectrum based on the ratio between the input spectrum smoothed by averaging previously computed spectral power and its minimum. Cohen presented the improved MCRA (IMCRA) algorithm in [Cohen, 03]; this improved algorithm detects noise-only regions based on the presence of speech probability, without hard distinctions between the absence or presence of speech. The MCRA and IMCRA algorithms have been used widely for noise estimation [Jeon, 11, Park, 12, Song, 12] because these algorithms perform well notwithstanding their simplicity.

In general, conventional algorithms for noise estimation and reduction have an important drawback. The noise in consecutive frames is correlated, similar to the speech signal, and this correlation depends on the type of noise. However, conventional algorithms that use smoothing do not consider this difference; rather, the use fixed smoothing parameters. To incorporate the noise type dependence of noise correlation into noise estimation and noise reduction algorithms, attempts have been made to optimize the smoothing parameters of noise estimation and noise reduction algorithms [Song, 12, Choi, 12, Yuan, 15]. Such parameter optimization requires to calculate the quality of speech metrics for all possible combinations of all parameters; then, a combination that maximizes the speech quality is considered to be the optimal one. These methods yield better speech quality than conventional algorithms that do not use optimized parameters. However, such parameter optimization is time-consuming, because speech quality has to be evaluated for all possible combinations of parameter values, and the number of such combinations can be very high. Consequently, significant research effort has been made to reduce the number of relevant parameters to three.

In this paper, a method for optimizing five major parameters for noise estimation is proposed using gradient descent, which is among the most widely used optimization methods in machine learning. The learning rate of gradient descent was set to a negative value for parameter optimization. After parameter optimization by gradient descent, the optimized parameter values were re-checked in a wider range of values, to prevent convergence to a local minimum.

2 Noise Estimation Algorithm

In this section, the improved minima controlled recursive averaging (IMCRA) algorithm is reviewed. This algorithm is one of the most widely used noise estimation algorithms, and it is used in this paper for parameter optimization. A common noise estimation technique is to recursively average previously computed noise spectral power during periods of speech absence, and hold the estimate during speech presence. Owing to the uncertainty regarding the presence of speech, in the IMCRA algorithm, the noise power is estimated by recursive averaging using the speech presence probability, as follows:

$$\bar{\lambda}_d(k,l+1) = \tilde{\alpha}_d(k,l)\bar{\lambda}_d(k,l) + (1-\tilde{\alpha}_d(k,l))|Y(k,l)|^2 \quad (1)$$

where $Y(k,l)$ is k-th frequency bin in l-th frame, and

$$\tilde{\alpha}_d(k,l) = \alpha_d + (1-\alpha_d)p(k,l) \quad (2)$$

is a time-varying frequency-dependent smoothing parameter adjusted by a fixed smoothing parameter α_d($0<\alpha_d<1$) and the conditional speech presence probability $p(k,l)$, which is estimated based on the noisy measurement.

The conditional speech presence probability given the *a posteriori* SNR is estimated as

$$\begin{aligned} p(k,l) &= \mathcal{P}\left(H_1(k,l)\middle|\gamma(k,l)\right) \\ &= \left\{1+\frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))exp(-\upsilon(k,l))\right\}^{-1} \end{aligned} \quad (3)$$

where γ and ξ represent the *a posteriori* SNR and the *a priori* SNR respectively, $\upsilon = \gamma\xi/(1+\xi)$, and $q(k,l)$ is the *a priori* probability for speech absence.

To calculate $q(k,l)$, two-step voice activity detection (VAD) is undertaken, using a hard decision and a soft decision, respectively. In the first step of VAD, to detect the speech activity roughly, the noise spectrum is smoothed in the frequency and time domains, respectively, as follows:

$$S_f(k,l) = \sum_{i=-w}^{w} b(i)|Y(k-i,l)|^2 \quad (4)$$

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)S_f(k,l) \quad (5)$$

where α_s($0<\alpha_s<1$) is a smoothing parameter, $b(i)$ is a Hanning window of length 2w+1 and $S_f(k,l)$ is a spectrum smoothed in the frequency domain by the convolution of the Hanning window and the noise spectrum.

The first step of VAD using the smoothed spectrum can be described by

$$I(k,l) = \begin{cases} 1 & if\ \left(\frac{|Y(k,l)|^2}{B_{min}S_{min}(k,l)} < \gamma_0\right)\ and\ \left(\frac{S(k,l)}{B_{min}S_{min}(k,l)} < \zeta_0\right) \\ 0 & otherwise \end{cases} \quad (6)$$

where $I(k,l)$ is the absence of speech identifier, γ_0 and ζ_0 are thresholds for identifying the noise components, B_{min} is a bias for compensating the minimum that is proportional to the fluctuation of noise, and

$$S_{min}(k,l) = min\{S(k,l')|l-D+1 \leqq l' \leqq l\} \quad (7)$$

is the minimum of the smoothed spectrum in a finite window of length *D*.

Based on the result of first VAD, $I(k,l)$, smoothing is performed with noise components in the second step of VAD.

$$\tilde{S}_f(k,l) = \begin{cases} \frac{\sum_{i=-w}^{w} b(i)I(k-i,l)|Y(k-i,l)|^2}{\sum_{i=-w}^{w} b(i)I(k-i,l)} & if\ \sum_{i=-w}^{w} b(i)I(k-i,l) \neq 0 \\ \tilde{S}(k,l-1) & otherwise \end{cases} \quad (8)$$

$$\tilde{S}(k,l) = \alpha_s \tilde{S}(k,l-1) + (1-\alpha_s)\tilde{S}_f(k,l) \quad (9)$$

where α_s and $b(i)$ have the same values as in Equations (4) and (5).

Based on the smoothed noise components, the *a priori* probability for the absence of speech, $q(k,l)$, can be calculated using a soft decision, as follows:

$$q(k,l) = \begin{cases} 1 & if\ (\tilde{\gamma}_{min}(k,l) \leq 1)\ and\ (\tilde{\zeta}(k,l) < \zeta_0) \\ \frac{\gamma_1 - \tilde{\gamma}_{min}(k,l)}{\gamma_1 - 1} & if\ (1 < \tilde{\gamma}_{min}(k,l) < \gamma_1)\ and\ (\tilde{\zeta}(k,l) < \zeta_0) \\ 0 & otherwise \end{cases} \quad (10)$$

where γ_1 is a threshold, and $\tilde{\gamma}_{min}(k,l)$ and $\tilde{\zeta}(k,l)$ are instantaneous SNRs calculated by

$$\tilde{\gamma}_{min}(k,l) = \frac{|Y(k,l)|^2}{B_{min}\tilde{S}_{min}(k,l)}, \qquad \tilde{\zeta}(k,l) = \frac{S(k,l)}{B_{min}\tilde{S}_{min}(k,l)} \quad (11)$$

where $\tilde{S}_{min}(k,l)$ is the minimum of $\tilde{S}(k,l)$ in a local window, and B_{min} is a bias for compensating the minimum.

The values of $\gamma(k,l)$ and $\xi(k,l)$ are estimated as

$$\gamma(k,l) \approx |Y(k,l)|^2 / \bar{\lambda}_d(k,l-1) \quad (12)$$

$$\xi(k,l) = \alpha G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha)max\{\gamma(k,l)-1, 0\} \quad (13)$$

where α(0< α<1) is a weighting factor that controls the tradeoff between the reduction of noise and speech distortion, and G_{H_1} is the spectral gain during the presence of speech.

In this paper, we optimize five parameters for noise estimation; α_d in Equation (2), B_{min} in Equations (6) and (11), α in Equation (13), ζ_0 in Equations (6) and (10), and α_s in Equations (5) and (9). These parameters are commonly considered to be affected by the noise characteristics.

3 Parameter Optimization

As mentioned in the introduction, noise is correlated in consecutive frames; thus, smoothing parameters for noise estimation and noise reduction algorithms should be

properly optimized [Choi, 12, Song, 12, Yuan, 15] for various noise scenarios. However, conventional methods that evaluate the speech quality using all combinations of parameters are too time-consuming for parameter optimization. Thus, in this section, we propose a method for parameter optimization that maximizes the quality of enhanced speech in various noise environments using a gradient descent algorithm.

The gradient decent algorithm is one of the optimization algorithms that minimize the error by updating parameter values in the opposite direction to the error gradient, as described below:

$$P_{n+1} = P_n - \gamma \nabla F(d) \tag{14}$$

where P_n is the set of parameter values in n-th iteration, γ is the learning rate, and $\nabla F(d)$ is the gradient of the (d-dimensional error) function for which the optimization is sought. As shown in this equation, if the gradient of the function is known, the set of parameter values can be updated in the direction of decreasing error, as long as the gradient is not zero. Unfortunately, the gradient descent method does not always find the optimal solution. The optimization process terminates at a point for which all gradient components are zero, corresponding to a minimum of the optimized function. However, this point may be a local minimum, rather than the global minimum; in this case, the found set of parameters is sub-optimal. To alleviate this sub-optimality problem, the optimization process is usually repeated many times with different initial values.

Unfortunately, the function that we seek to optimize in this paper is too complicated to calculate its gradient because parameters are not independent. Thus, we estimated the gradient based on the difference between a specific point and its neighbor point. In addition, we updated only the parameter that yielded the steepest gradient because the speech quality dependence on parameters is not convex and can be decreased by small changes of parameter values. To update only one parameter that yields the steepest gradient, the dimension in D-dimensional parameter vector that corresponds to the steepest gradient was defined by

$$I = maxidx\{Q(\tilde{P}_{n,i}) - Q(P_n),\ i = [0, 1, \dots, D]\} \tag{15}$$

where P_n is the initial vector of parameter values during n-th iteration, $\tilde{P}_{n,i}$ is the neighbor vector of parameter values that differs from the initial vector only in i-th dimension, and Q is the quality of enhanced speech by the algorithm in section 2 for a given vector of parameter values, evaluated using a composite measure corresponding to a combination of the distance measure (IS), the perceptual evaluation of speech quality measure, the log-likelihood ratio, the Itakura–Saito distance measure (IS), and the weighted-slope spectral distance; the combination was built as follows [Hu, 08]:

$$C_{ovl} = 0.279 - 0.011 \cdot IS + 1.137 \cdot PESQ + 0.041 \cdot LLR - 0.008 \cdot WSS \tag{16}$$

If the steepest gradient is determined for the dimension d, the gradient component for d-th dimension is calculated; otherwise, the gradient is set to 0.

$$\nabla F(d) = \begin{cases} Q(\tilde{P}_{n,I}) - Q(P_n) & if\ I = d \\ 0 & otherwise \end{cases} \tag{17}$$

Here, $\tilde{P}_{n,I}$ is the neighbor set of parameter values within the maximal gradient, and the gradient is limited by a maximum (=0.03) and minimum (=0.01).

As mentioned above, the objective is to determine the set of parameters that maximizes the quality of speech for a given noise scenario. Thus, the learning rate γ in Equation (14) was set to a negative value, to cast gradient descent as gradient ascent. Before iterations for parameter optimization by gradient descent, we selected initial parameters based on a rough speech quality evaluation using large parameter value steps. Because the set of parameters to be optimized contains three smoothing parameters, and noise power is estimated assuming that noise changes slower than speech, the values of the smoothing parameters were limited to [0.50, 0.99] and the other two parameters were 1.7 times found in the range [0.50, 0.99]. In steps of 0.1, the speech quality for all combinations of values [0.6, 0.7, 0.8, 0.9] was evaluated; overall, 4^D evaluations were required for determining the initial set of parameter values. After the initialization, parameter optimization was performed using Equations (15) and (17), in the step of 0.01. In this procedure, the quality of speech was evaluated for neighbor parameter sets, within a three-step distance in each dimension from the initial parameter set. Figure 1 shows an example of a neighbor parameter set, for two-dimensional parameter vectors and for the initial parameter set [a, b].

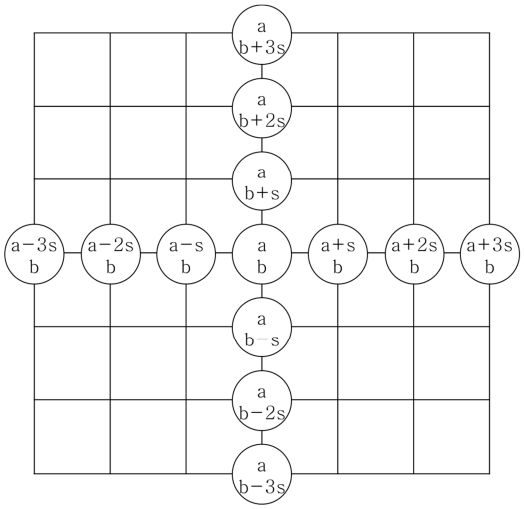

Figure 1: Example of a neighbor parameter set when the initial parameter set is [a, b].

Using this method, 6D+1 speech quality evaluations were performed in each iteration, when the dimensionality of the parameter set was D. After detecting the optimal parameter set, we re-checked that the detected parameter set is globally optimal, using larger steps (=0.04~0.06), to avoid the local minimum problem.

4 Performance Evaluation

In this section, the performance of the proposed method is described. Because the purpose of this study is to efficiently optimize noise estimation parameters, we measured the required overall number of speech quality evaluations. If the optimal parameter set was determined correctly, the quality of speech was improved by the optimal parameter set. Thus, we measured the speech quality obtained using the optimal parameter set.

To evaluate the method's performance, we considered four types of noise (babble, destroyer engine, Volvo, and white) from NoiseX-92, and 20 speech recordings (1 male and 1 female talker) from the TIMIT database. Each noise and speech samples were mixed by 0-, 5-, 10-, and 15-dB SNR. Two types of speech per each talker in 0-dB conditions, with the strongest noise components, were used for parameter optimization using noisy speech, and eight types of speech per each talker were used for validation of the calculated optimal parameter value sets.

Figure 2 shows the proposed extraction method for optimization of two parameters (1st parameter is α_d, and 2nd parameter is B_{min}). In this figure, the mesh plot shows the speech quality for all combinations of the two parameters, the gray filled diamonds show the optimal parameters, the gray filled circles are the points for determining the suitable initial point, and the black crosses are the optimal points (values) found by gradient descent in every iteration. As shown in this figure, the proposed extraction method can achieve parameter optimization in a few iterations after the initial set of parameters is determined.

In this paper, five parameters; α_d in Equation (2), B_{min} in Equations (6) and (11), α in Equation (13), ζ_0 in Equations (6) and (10), and α_s in Equations (5) and (9) were optimized, and the number of speech quality evaluations for parameter optimization were 1274, 1350, 1298, and 1328 for different noise scenarios, with the average of 1312.5 iterations. On the other hand, parameter optimization using conventional methods requires 312,500,000 $((0.99\text{-}0.50)/0.01+1)^5$) evaluations for one noise scenario. This suggests that the proposed method optimizes using 0.00042% evaluations compared with conventional methods, when five parameters are optimized.

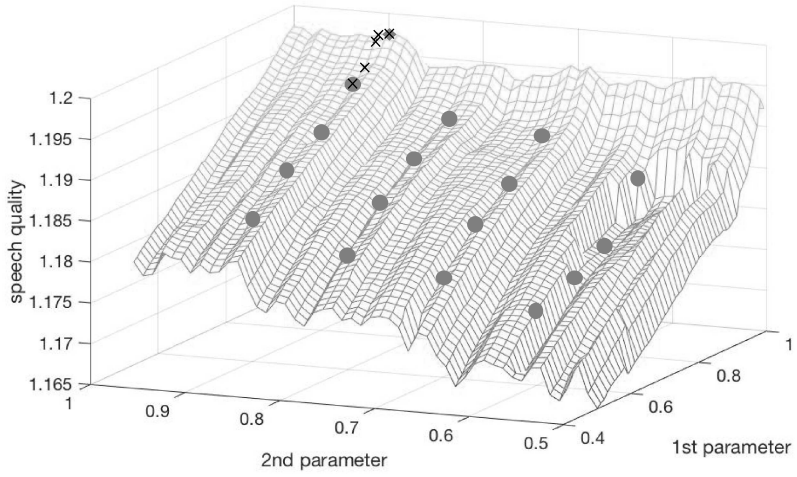

Figure 2: Proposed parameter optimization method, showing the optimal parameters (gray filled diamonds), parameter combinations for finding the initial point (gray filled circles), and the optimal points found by gradient descent in every iteration (black crosses).

Table 1 shows the optimal sets of parameters extracted using the proposed method, for different noise scenarios. The first parameter, α_d, used for estimating the noise, is lower for babble noise, which was the most unstable noise in this study. This trend is similar to that in other studies that used the IMCRA noise estimation algorithm [Song, 12 and Yuan, 15]. The second parameter used for compensating the minimum tended to be smaller for stable noise. This means that the minimum was similar to the estimated noise power, and the fluctuation of power was weaker.

	α_d	B_{min}	α	ζ_0	α_s
babble	0.57	1.666	0.96	1.666	0.98
destroyer engine	0.96	1.530	0.96	1.394	0.92
Volvo	0.98	1.122	0.97	1.292	0.95
white	0.95	1.479	0.96	1.394	0.90

Table 1: Optimal parameters for different noise scenarios

The third parameter for estimating the *a priori* SNR was similar across all noise conditions. The fourth parameter for estimating the VAD was smaller for stable noise.

This suggests that speech components can be easily detected. The last parameter for smoothing the input spectrum was smaller for stable noise, similar to previous findings [Yuan, 15].

Within these optimal parameter sets, we estimated the noise using the IMCRA algorithm and calculated the spectral gain based on LogMMSE [Ephraim, 85], then enhanced the speech from noisy speech by multiplying the spectral gain by the spectrum of noise. Table 2 shows the average quality of enhanced speech obtained using conventional parameter sets and optimized parameter sets, for different noise scenarios. The quality of enhanced speech was measured using Equation (16), and eight types of speech that were not used in the parameter optimization study were used for evaluating the speech quality. As shown in Table 2, the quality of the enhanced speech using optimized parameters was higher than that obtained using conventional parameters, for all conditions, and increased on average by 1.1307%, 3.097%, 3.742%, and 3.861% for 0-, 5-, 10-, and 15-dB SNR, respectively. The babble noise, which was the most unstable noise in this study, showed a noticeable increase in the speech quality for the optimized parameter set. The speech quality for the babble noise increased by 1.607%, 7.876%, 11.986%, and 13.095% for 0-, 5-, 10-, and 15-dB SNR, respectively. This suggests that the optimized parameters are more effective for unstable noise.

SNR	0 dB		5 dB		10 dB		15 dB	
parameter	conv	opt	conv	opt	conv	opt	conv	opt
babble	1.307	1.328	1.550	1.672	1.910	2.139	2.408	2.724
destroyer engine	1.662	1.669	2.108	2.134	2.881	2.915	3.480	3.507
Volvo	1.707	1.735	2.200	2.254	2.966	3.004	3.517	3.555
white	1.638	1.652	1.844	1.858	2.237	2.248	2.699	2.712

Table 2: Composite overall speech quality with conventional parameter set and optimal parameter set

5 Conclusions

In previous studies by other researchers, the number of parameters for noise evaluation was set to three. This was because the computational burden for systems with more than three parameters was too large because the amount of computation increases exponentially with the number of parameters. For three parameters to be optimized in 0.01 steps with each parameter value in the [0.8, 0.99] range, conventional methods that rely on all combinations of parameters require 8000 speech

quality evaluations for determining the optimal combination. Thus, conventional methods are inefficient for parameter optimization.

In this study, we proposed an efficient method for the optimization of major parameters related to noise estimation. The total number of speech quality evaluations for optimization using the proposed method was reduced by 99.99958% compared with conventional methods when five parameters were optimized. Thus, 10 s of computation time were required for one speech quality evaluation, and an average of 1312.5 speech quality evaluations for one noise scenario described in Section 4 took 3.646 h for the optimization of five parameters using the proposed method. The extracted optimal parameter values increased the speech quality on average by 1.1307%, 3.097%, 3.742%, and 3.861% for 0-, 5-, 10-, and 15-dB SNR, respectively.

6 Future Work

Although the proposed optimization method dramatically reduces the total number of calculations, a disadvantage of this algorithm is that it must be run offline and must be modified or recalculated after a new noise scenario is added. In the future, modified noise estimation algorithms will be developed that can automatically adjust the values of parameters according to the noise characteristics.

Acknowledgements

This research was supported by the KERI Primary Research Program through the National Research Council of Science & Technology (NST) funded by the Ministry of Science and ICT (MSIT) (No. 18-12-N0101-39) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT(NRF-2016R1A2B4015370)

References

[Boll, 79] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." IEEE Transactions on acoustics, speech, and signal processing 27.2 (1979): 113-120.

[Cohen, 02] Cohen, Israel, and Baruch Berdugo. "Noise estimation by minima controlled recursive averaging for robust speech enhancement." IEEE signal processing letters 9.1 (2002): 12-15.

[Cohen, 03] Cohen, Israel. "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging." IEEE Transactions on speech and audio processing 11.5 (2003): 466-475.2003;11(5):466-75.

[Choi, 12] Choi, Jae-Hun, and Joon-Hyuk Chang. "On using acoustic environment classification for statistical model-based speech enhancement." Speech Communication 54.3 (2012): 477-490.

[Ephraim, 85] Ephraim, Yariv, and David Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator." IEEE Transactions on Acoustics, Speech, and Signal Processing 33.2 (1985): 443-445.

[Hu, 08] Hu, Yi, and Philipos C. Loizou. "Evaluation of objective quality measures for speech enhancement." IEEE Transactions on audio, speech, and language processing 16.1 (2008): 229-238.

[Jeon, 11] Jeon, Yu-yong, and Sang-min Lee. "A speech enhancement algorithm to reduce noise and compensate for partial masking effect." Journal of Central South University of Technology 18.4 (2011): 1121-1127.

[Martin, 01] Martin, Rainer. "Noise power spectral density estimation based on optimal smoothing and minimum statistics." IEEE Transactions on speech and audio processing 9.5 (2001): 504-512.

[Park, 12] Park, Yun-Sik, Gyu-Seok Park, and Sang-Min Lee. "Speech Enhancement Based on Modified IMCRA Using Spectral Minima Tracking with Weighted Subband Selection." Journal of the Institute of Electronics Engineers of Korea SP 49.3 (2012): 89-97.

[Song, 12] Song, Ji-Hyun, et al. "Speech Enhancement Based on IMCRA Incorporating noise classification algorithm." The Transactions of The Korean Institute of Electrical Engineers 61.12 (2012): 1920-1925.

[Yuan, 15] Yuan, Wenhao, and Bin Xia. "A speech enhancement approach based on noise classification." Applied Acoustics, Vol.96 (2015): 11-19, DOI:10.1016/j.apacoust.2015.03.005