**TELKOMNIKA**, Vol.14, No.1, March 2016, pp. 181~186
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i1.2021


# A Novel Scheme of Speech Enhancement using Power Spectral Subtraction - Multi-Layer Perceptron Network

**Budiman P.A. Rohman*, Ken Paramayudha, Asep Yudi Hercuadi**
Research Center for Electronics and Telecommunications, Indonesian Institute of Sciences
Kampus LIPI Gd.20 Lt. 4 Jl. Sangkuriang Bandung 40135, Indonesia
*Corresponding author, email: budiman.par@gmail.com

***Abstract***

*A novel method for eliminating noise from a noised speech signal in order to improve its quality using combined power spectral subtraction and multi-layer perceptron network is presented in this paper. Firstly, the contaminated speech signal was processed by spectral subtraction to enhance the clean speech signal. Then, the signal was processed by a neural network using the spectral subtraction parameters and result of estimated speech signal in order to improve its signal quality and intelligibility. The artificial neural network used was multi-layer perceptron network consisted of three layers with six input and one output. The neural network was trained with three speech signals contaminated with two level white gaussian noises in SNR including 0 dB and 30dB. The designed speech enhancement was examined with ten noised speech signals. Based on the experiments, the improvement of signal quality SNR was up to 7 dB when the signal quality input was 0dB. Then, based on the PESQ score, the proposed method can improve up to 0.4 from its origin value. Those experiment results show that the proposed method is capable to improve both the signal quality and intelligibility better than the original power spectral subtraction.*

***Keywords***: *speech enhancement, spectral subtraction, artificial neural network, multi-layer perceptron.*



## 1. Introduction

The speech enhancement is an important tool for supporting many applications especially in the telecommunication areas such as in the mobile communication. Others, this capability has direct influences to the performance of the human-machine interface applications such as speech recognition and speaker recognition which are very popular currently. In many situations, the high level noise degrading speech signal can decrease the performance of those applications. Moreover, it degrades the speech quality and intelligibility, and also affects negatively to the listener's perception especially in mobile communication cases [1]. This is the main problem of the speech enhancement which almost all of these speech enhancement approaches rely on the estimation of a short-time spectral gain [2].

There are several types of speech enhancement algorithm investigated by researchers over the world which can be separated to two groups e.g. single channel and multichannel speech method. Based on several researches, the multi-channel speech enhancement has better performance than single channel methods. However, because of its simplicity and low cost implementation, the single channel method is still worthy to be explored and improved. The most popular single channel speech enhancement method is spectral subtraction.

Spectral subtraction is well known noise reduction method and it is one of the first algorithms for speech enhancement facing musical noise. Firstly, it was investigated by Boll in 1979 [3]. In its first method, spectral subtraction is purposed for eliminating musical noise. In the same year, the spectral subtraction was improved by Berouti in 1979 [4]. He developed an over-subtraction constant for over-estimating the noised speech signal. Although the development of spectral subtraction has been begun since 1979, this method has been used in many applications until now because it is relatively inexpensive in computation [5]. However, spectral subtraction suffers from a problem of introducing artifacts like noise while removing residual noise. It then will influence to the both quality and intelligibility of estimated speech signal. Hence, current researches on spectral subtraction techniques are concentrated on decreasing or removing this noise. [6]

*Received May 13, 2014; Revised November 27, 2015; Accepted December 17, 2015*

Over these years, spectral subtraction has been modified and improved by many researchers over the world. In 2002, Sunil Kamath [7] investigated the multi band spectral subtraction for eliminating the colored noise in corrupted speech signal. This research provides the definite improvement over the conventional power spectral subtraction. Othres, an investigation of an improved spectral subtraction using perceptual weighting filter was held by RM Udrea [8]. In this research, he had improved the quality of speech signals. Then, in 2011, Verteletskaya [9] proposed the modified spectral subtraction for eliminating residual noises. Also, the speech enhancement using spectral subtraction in wavelet domain was proposed by Nishimura et.al in 1998 [10].

In this paper, we proposed the novel approach of speech enhancement by combining the spectral subtraction and artificial neural network for optimizing the speech enhancement. The neural network used is multi-layer perceptron network because among other network structure this network is most successful design [11]. In this scheme, spectral subtraction is used as a main process and the neural network became an optimizer by recalculating and then improving the quality and intelligibility of signal output of spectral subtraction. The purpose of this research is that by using this method, speech signal will be enhanced better than origin power spectral subtraction method but still keep its simplicity in the computation.

## 2. Spectral Subtraction and Artificial Neural Network

### 2.1. Spectral Subtraction (SS)

Spectral Subtraction is a speech enhancement method which process in the frequency domain. In this algorithm, there are two steps i.e. VAD (Voice Activity Detection) step and spectral subtraction. In the step of VAD, the speech signal is processed for labeling whether the framed signal is voice, unvoiced or silent signal. This will lead to next spectral subtraction step as a primary step of this speech enhancement method.

Assume y(n) = x(n) + d(n) is the sampled noisy speech signal consisting of the clean signal x(n) and the noise signal d(n). The frequency domain of signal in $k^{th}$ frame can be represented as below,

$$Y(\omega, K) = DFT(w(n,k)) \tag{1}$$

$$Y(\omega, K) = X(\omega, k) + D(\omega, k) \tag{2}$$

After speech signal y(n) is transformed into the frequency domain, the spectrum, mean and standard deviation of first framed signal k=1 is considered as noise.

$$No(\omega) = Y(\omega, k) \text{ , k=1} \tag{3}$$

$$\phi = 20\{\log 10(Y(\omega,k)) - \log 10(No(\omega))\} \tag{4}$$

$$\phi' = \frac{1}{L}\sum \phi \tag{5}$$

$\phi$ is magnitude spectral distance between signal and noise and $L$ is the length of frame. $\phi'$ is the mean of magnitude spectral distance in a frame. This value then will be compared with the predetermined thresholds of noise margin ($N_m$) and hangover ($h$). Commonly, the value used for those two thresholds are 3 and 8 respectively. If the mean of magnitude spectral distance is lower than the noise margin, this frame will be labeled as noise signal. In contrast, if the mean is higher than the hangover constant, this frame will be considered as speech signal.

In the spectral subtraction step, this research uses factor of over-subtraction and spectral-floor based on posteriori SNR.

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} \left|\hat{Y}(\omega)\right|^2 - \alpha\left|\hat{D}(\omega)\right|^2 & jika \left|\hat{Y}(\omega)\right|^2 > \left|\hat{D}(\omega)\right|^2 \\ \beta\left|\hat{D}(\omega)\right|^2 & others \end{cases} \tag{6}$$

α is an over subtraction factor for overestimating the noise spectrum. β is a spectral floor factor which is needed to avoid the elimination of speech signal at the lowest level. The optimal range of β is between 0.1 and 0.001. Below is the equation used for calculating the over-subtraction factor.

$$\alpha = \alpha_0 - \frac{3}{20} SNR_{posterior} \qquad \text{-5 dB} \leq SNR_{posterior} \leq \text{20 dB} \tag{7}$$

Then, the SNR (Signal to Noise ratio) is calculated using below equation.

$$SNR_{posteriori}(\omega) = \frac{|Y(\omega)|^2}{\left|\hat{D}(\omega)\right|^2} \tag{8}$$

Where $\alpha_0$ is the targeted α when signal in 0dB quality. Then, for power spectral subtraction step, the optimal range from $\alpha_0$ is in between 3 and 6.

### 2.2. Artificial Neural Network (ANN)

Artificial Neural network is designed based on the biological human brain neuron construction. As a human brain representation, the neural network generally consists of neuron, weight, activation function and layer. Neuron is a simple processing unit. In this part, the multiplication of weight and activation function is processed. The weight is the weight value of input in neural network. This value will be adapted in the training process. Activation Function is needed for a threshold process after summing the weighted input. Layer is a set of neurons in the neural network [11].

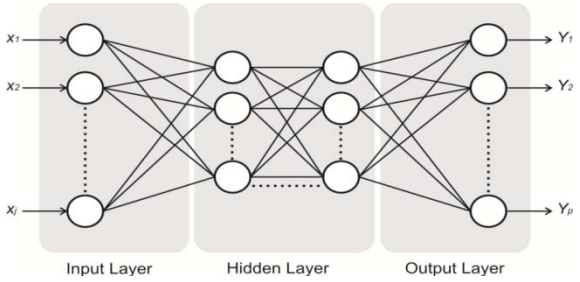


Figure 1. Common structure of artificial neural network

See figure on the top, $x_j$ is input of neural network. The summation of weighted input, $v_k$, can be computed as:

$$v_k = \sum_{j=1}^{p} w_{kj} x_j \tag{9}$$

The output of the neuron, $y_k$, would therefore be the outcome of some selected activation function on the value of $v_k$.

In the Multi-Layer Perceptron structure, the neural network consists of several hidden layers between input and output layer [11]. Generally, there are weights in between neuron in adjoin layer. This network is capable to solve complex pattern. However, the training and computation of this network is more complex than the single layer network structure.

## 3. Research Method

In this research, we proposed the use of Multi-Layer Perceptron Network (MLPN) with three hidden layers using a logarithmic sigmoid activation function. Neural network consisted of

6 inputs, 1 output and 3 layers with 8, 4 and 2 neurons for each layer respectively. Each neuron had a bias value. Inputs of neural network consisted of enhanced speech signal, estimated noise, mean of estimated noise, estimated SNR, gradient of estimated SNR and VAD flag. Output of neural network was the estimated clean speech signal.

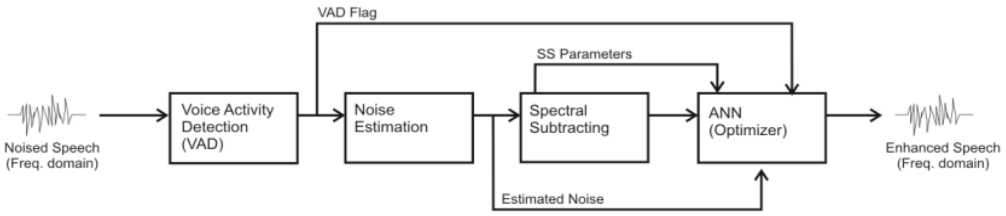


Figure 2. General design of proposed speech enhancement

After sampling, frame-blocking and windowing Hamming (with time length is 25ms and 40% overlapping) processes, noised speech signal was transformed into frequency domain using Discrete Fourier Transform. Then, the signal in frequency domain was processed using Spectral Subtraction which was separated into two steps i.e. VAD and Spectral Subtraction. VAD used magnitude spectral distance with noise margin=3 and hangover constant= 8. In the spectral subtraction steps used parameter $\beta = 0.03$. After spectral subtraction process, the next step is neural network which lead to better quality speech signal.

Training of neural network in this filter was configured by learning rate 0.98, maximum epoch 1000 and target error Mean Squared Error (MSE) $1x10^{-1}$. Training algorithm used in this experiment was *Lavendberg Marquadt*. The initial weights of the neural network were selected randomly. The activation function used in this neural network was logarithmic sigmoid function. For training step, the filter was trained by 3 different noised speeches with 2 noise levels SNR i.e. 0dB and 30dB. Noise type used in this research was white noise. The target of neural network training was a clean speech signal.

The process of enhancement will be run sequentially starting with spectral subtraction and then neural network. Firstly, the contaminated speech signal is processed by spectral subtraction to get the first clean speech estimation. Then, this estimated signal will be processed further by neural network. The function of neural network in this method was for improving the quality and intelligibility of estimated speech signal after spectral subtraction. Using the used spectral subtraction parameters and estimated noise, this network was trained for reconstructing the clean speech signal (See Figure 2).

The analysis of this method (titled as NN-SS) could be divided to three including time domain analysis, frequency domain analysis and comparison of signal quality level (SNR) and PESQ (Perceptual Evaluation of Speech Quality) score. PESQ is the most complex to compute and this is the one recommended by ITU-T for speech quality assessment of 3.2 kHz handset telephony and narrow-band speech codecs [12]. PESQ measures performed modestly well in terms of predicting both quality and intelligibility [13]. The score of PESQ is ranged from 0-4.5 which the high score means high quality and intelligibility. All of those analysis will be compared to the original power spectral subtraction method (titled as SS).

## 4. Simulation Result and Analysis

The training process of designed neural network in the speech enhancement was converged before the goal was achieved. The training stopped at 87 epochs with Mean Square Error (MSE) was around 0.8755. Next, the trained algorithm was tested by degraded noised for further analysis.

### 4.1. Time Domain Analysis

Below the comparison of speech signal using spectral subtraction and improved spectral subtraction using neural network.

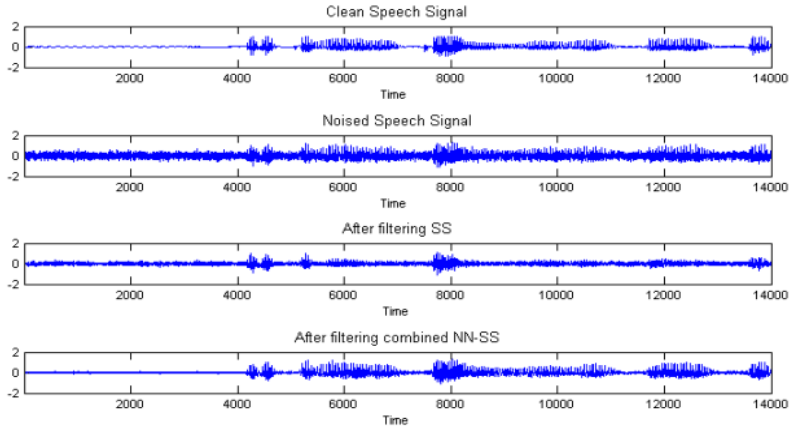


Figure 3. Comparison time domain results of NN-SS and SS with input SNR 0 dB

Figure 3 shows the comparison of time domain signal among clean speech signal, noised speech signal with SNR 0 dB, estimated clean speech signal by SS and estimated speech signal by NN-SS. It can be seen clearly that the quality of the signal after the NN-SS was improved better than by SS only. Moreover, by the NN-SS, the clean speech signal could be reconstructed and optimized. In details, for example, in the sample number 1 to 4000 the signal which was contained by background noises only was successfully suppressed to its very minimum value. Then, in the sample number around 5000 to 7000, NN-SS successfully reconstructed and resembled the original waveform which had been distorted by the SS. Others, NN-SS could repair the speech signal which actually had been eliminated by SS. This ability could be seen in the sample number around 9000 to 13000.

### 4.2. Frequency Domain Analysis

Figure 4 shows the filtering result by both NN-SS and SS in the spectrogram. After SS, there was still a noise signal left spreading around the original signal. By NN-SS, this noise had been eliminated without destructing the original form of speech signal resulted by the spectral subtraction filter. In several samples, NN-SS reconstructed the speech signals which actually had been eliminated by SS. Overall, based on those frequency analyses, NN-SS had result better signal quality than SS as well.

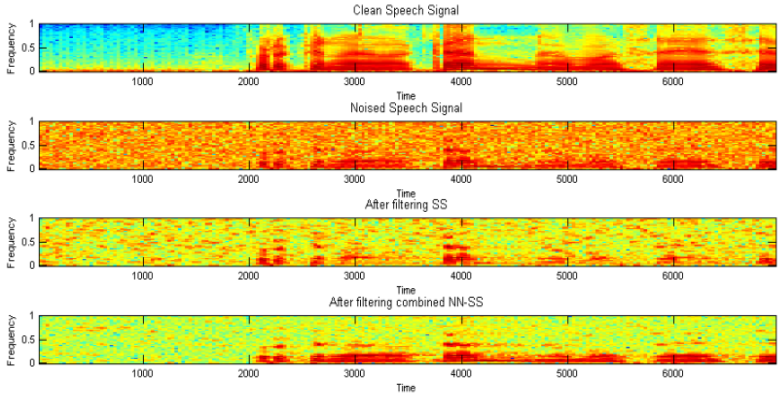


Figure 4. Comparison of spectrogram of result by NN-SS and SS with input SNR 0dB

### 4.3. Comparison of SNR and PESQ Score

In addition, the NN-SS was examined based on its signal quality and intelligibility improvement. Below is the mean result score of testing of proposed speech enhancement by 10 degraded speech signals within SNR ranged from 0 to 15dB,

Table 1. Comparison SNR of results of NN-SS and SS

| No | SNR | | | PESQ | | |
|---|---|---|---|---|---|---|
| | Input | SS | NN-SS | Input | SS | NN-SS |
| 1 | 0 | 3.12 | 7.01 | 1.56 | 1.56 | 1.93 |
| 2 | 5 | 5.16 | 11.03 | 1.92 | 2.02 | 2.42 |
| 3 | 10 | 9.31 | 12.84 | 2.31 | 2.48 | 2.63 |
| 4 | 15 | 14.53 | 13.77 | 2.67 | 2.94 | 2.84 |

Based on the SNR value, the table shows that the NN-SS could improve the signal quality up to 7 dB when input signal SNR was 0dB. The improvement decreased steeply follow the quality of the input signal. For example, when the input signal SNR was 5dB, the improvement was approximately 6dB. However, those improvements were higher than the SS algorithm especially when the signal input quality was very low (0-10dB). Furthermore, depend on the PESQ scores; the output signal of NN-SS had the higher score than the SS by up to 0.4. It means that the NN-SS has better performance than the SS in both quality and intelligibility.

## 5. Conclusion

This paper has presented the results of application of speech enhancement using combination of power spectral subtraction and multi-layer perceptron network, namely Neural Network-Spectral Subtraction (NN-SS) in the presence of white Gaussian noises. Overall, NN-SS is capable to enhance the speech signal relatively better than the origin SS. Furthermore, in the low signal quality input, this method has significant improvement. The future researches can be conducted in the case of in the presence of the non-stationary noises. Others, the further investigation can be held leads to minimize the complexity of computation, especially if it will be implemented in real application.

## References

[1] Yan, Wang Guang, Geng Yan Xiang, and Zhao Xiao Qun. A Signal Subspace Speech Enhancement Method for Various Noises. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(2): 726-735.
[2] Ou Shifeng, Chao Geng, Ying Gao. Improved a Priori SNR Estimation for Speech Enhancement Incorporating Speech Distortion Component. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11(9): 5359-5364.
[3] Boll Steven F. *A spectral subtraction algorithm for suppression of acoustic noise in speech*. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.1079; 4.
[4] Berouti M, R Schwartz, John Makhoul. *Enhancement of speech corrupted by acoustic noise*. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. 1979; 4.
[5] Vaseghi Saeed V. Advanced digital signal processing and noise reduction. John Wiley & Sons. 2008.
[6] Goel Paurav, Anil Garg. Review of Spectral Subtraction Techniques for Speech Enhancement 1. (2011).
[7] Kamath Sunil, Philipos Loizou. *A multi-band spectral subtraction method for enhancing speech corrupted by colored noise*. IEEE international conference on acoustics speech and signal processing. 2002; 4.
[8] Udrea Radu Mihnea, Nicolae D Vizireanu, Silviu Ciochina. An improved spectral subtraction method for speech enhancement using a perceptual weighting filter. *Digital Signal Processing.* 2008; 18(4): 581-587.
[9] Verteletskaya Ekaterina, Boris Simak. Noise reduction based on modified spectral subtraction method. *IAENG International journal of computer science.* 2011; 38(1): 82-88.
[10] Nishimura Ryouichi, et al. Speech enhancement using spectral subtraction with wavelet transform. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science).* 1998; 81(1): 24-31.
[11] Hu Yu Hen, Jenq-Neng Hwang. Handbook of neural network signal processing. CRC press. 2001.
[12] Hu Yi, Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on.* 2008; 16(1): 229-238.
[13] Ma Jianfen, Yi Hu, Philipos C Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America.* 2009; 125(5): 3387-3405.