# Analysis of Noise Reduction Techniques in Speech Recognition

Bo Zheng[1] ,Jinsong Hu*[1] , Ge Zhang [2],Yuling Wu[2], Jianshuang Deng[3]
1. South China University of Technology
2. System Operation Department (Power Dispatching and Controlling Center), Guangzhou Power Supply Co., Ltd.
3. Guangzhou Kinth Network Technology Co., Ltd.
Guangzhou, China
holy5pb@163.com, cshjs@scut.edu.cn, 717754917@qq.com, ansonwyl@qq.com, 173309406@qq.com

***Abstract*—Compared with keyboard input, speech recognition has its strengths——it is more in line with the natural communication of people and frees people's hands. In recent years, with the development of artificial intelligence, speech recognition technology has developed rapidly, and the recognition rate is generally in excess of 90%. However, in the case of environmental noise, recognition rates of current speech recognition products have been severely reduced, and most of them cannot work normally. How to enhance these voices with noise and restore the information of the original signal to the greatest extent is an urgent problem. This article summarizes some current mainstream noise reduction algorithms, and compares them by explaining principles of these algorithms. Finally, this article makes a brief prediction of the general development direction of the speech enhancement field in the future.**

***Keywords—speech recognition; noise reduction; speech enhancement; signal processing***

## I. INTRODUCTION

Speech recognition is a technology in which a machine converts a voice signal into corresponding text or commands. Automatic Speech Recognition (ASR) technology, which mainly completes the conversion from speech to text, belongs to non-specific person speech recognition. Up to now, speech recognition has changed many aspects of human life. From voice typewriters to voice control required for specific environments, it has brought a lot of convenience to people.

Speech recognition has incomparable advantages over keyboard input. Speech is the most frequent way of communication in daily life of people. It does not require any external tools and is simple and convenient. As a Human-Computer Interaction method close to the natural way of communication, voice input does not need to undergo special learning like keyboard input methods. The elderly, children, and people with disabilities can easily use it. It frees people's hands and brings great benefits for ordinary people as in many occasions, since, in some cases, it is inconvenient to input with your fingers, such as driving in a car.

Although people have started to study speech recognition a long time ago, until the last ten years, after the breakthrough progress of artificial intelligence, especially deep learning, the accuracy of speech recognition has been greatly improved, and speech recognition applications and products are beginning to come into the spotlight.

Some of the well-known speech recognition software products are listed as follows: Microsoft Cortana which is integrated in Windows System, Baidu Du Smart Speaker, Apple Siri, Xiaomi Xiao AI Speaker, etc. These companies and related technical teams claim that the recognition rates of these software products nearly break 95% or more. It seems that the speech recognition problem has been completely solved.

However, we tried to test the above software with multiple conversation recordings of the power dispatcher of China Southern Power Grid, including Microsoft Cortana, Google Voice Assistant, Baidu Du Smart Speaker, Apple Siri, and Xiaomi Xiao AI Speaker, but no one can make recognition rate break 20%, and the recognition result can hardly be understood by ordinary people. These recordings are all made indoors, and the ambient noise is so low that the human ear can recognize them easily. In order to confirm this problem, we also conducted repeated tests with conversation recordings in other indoor environments, and the results were similar, most of the software can not recognize them at present. Obviously, this situation is completely different from the high-accuracy speech recognition we know. We analyze that the reason is that the speech of the other person may be a kind of noise in some sense, which interferes with speech recognition. The current speech recognition software can only recognize one person's speech, and even requires people to pause a little after each speak sentence.

The current speech recognition technology still needs to be further improved. If it is not possible to recognize conversations in a weak noise environment, its application will be extremely limited. Many important applications that require speech recognition, such as a power intelligent dispatching voice response system, a large company's customer service response system, and humanoid robots, currently cannot apply speech recognition technology. Hence, the noise reduction technology in speech recognition has become the bottleneck of natural language processing and is the key to its future development.

The organizational structure of this paper is as follows: Section 2 introduces the classification of noise, Section 3 introduces the three basic categories and algorithms of speech enhancement, and explains some of the current mainstream noise reduction scheme principles; Section 4 lists the performance of some noise reduction algorithms evaluation method. In the last section, we made a brief summary and

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

made predictions for the development direction of speech noise reduction.

## II. CLASSIFICATION OF THE NOISE

During the transmission process, the voice signal will inevitably be affected by many factors such as the environment and the transmission medium itself, which will cause the quality (i.e. signal-to-noise ratio and intelligibility) of the pure voice signal to decrease. In the engineering field, speech noise reduction is an important precursor for a complete speech recognition system. The key step of speech recognition is to preprocess the speech signals. In order to effectively extract the features of the target speech, anti-noise and enhancement algorithms are needed to remove the interference signals. An effective anti-noise algorithm is a prerequisite for the system to correctly recognize speech signals. The sources of noise are very wide, usually including social life noise, traffic noise, construction work noise, industrial noise and so on. Noise is unpredictable, and characteristics are infinitely variable. For a variety of noise, we can classify the noise from multiple perspectives. Reference [1] gives a classification method:

(1) One is additive noise. Additive noise and speech signals satisfy the additive relationship. In our real life, additive noise can be seen everywhere, such as broadband noise, fan sounds, speech, car sounds, and impulsive noise. Some speech acquisition devices can be approximated as linear systems where the signal amplitude is directly proportional to their energy. Therefore, after being collected by the linear speech acquisition device, speech and additive noise are still additive. For a speech recognition system, the influence of additive noise on its recognition accuracy occupies a large proportion, so it is necessary to study the removal of additive noise.

(2) Other noise is multiplicative noise. When the noise and the speech signal are multiplicative in the frequency domain, this noise is called multiplication noise. This noise is also called convolution noise because they satisfy the convolution relationship because it is also called convolution noise. At this time, when the signal passes through the channel, a part of the frequency spectrum becomes weak. However, multiplicative and additive noise can be transformed by a homomorphic transformation.

Generally speaking, in the noise research, additive noise is generally more than multiplicative noise. The principle of speech noise reduction is to filter out the noise signal doped in the speech signal as much as possible, so as to restore the original state. However, in general, noise is unpredictable. Hence, the purpose of speech noise reduction is to eliminate as much noise interference as possible in the speech signal, thereby improving speech quality and intelligibility. Speech enhancement algorithms for noise reduction can be divided into three basic categories: filtering techniques, spectral reconstruction, and model-based methods [2].

## III. CLASSIFICATION AND PRINCIPLE OF SPEECH NOISE REDUCTION ALGORITHM

The research on noise reduction technology has a long history. Before speech recognition occurs, noise reduction was mainly used in music playback equipment and related places, such as radios, televisions, amplifiers, and theater equipment. Early noise reduction mainly relied on hardware Realization, such as Dolby noise reduction circuit, microphone noise reduction array, etc.

Microphone Noise Reduction has been widely used in mobile phones in recent years. At present, most mobile phones use dual microphone noise reduction. The principle of the Microphone Noise Reduction is that the ratio of the distance between the microphones at the upper and lower ends of the mobile phone and the speaker's mouth is relatively large, resulting in a large difference in the speaker's sound intensity received by the upper and lower microphones, and the distance ratio of the noise is relatively small, which can be used to eliminate noise, including the voices of other people in the environment—for speech recognition, the voice of non-target people is essentially noise. However, in a two-person conversation scenario (non-mobile phone conversation), the voice of the other person cannot be cancelled with this technology, and as described above, even if there is no other environmental noise at this time, the current speech recognition software still cannot recognize it.

Obviously, in order to realize the recognition of two people's conversational speech, we cannot rely on the current hardware to reduce noise, we must develop related speech noise reduction algorithms. Generalized speech noise reduction can also be called speech enhancement, and its general process is shown in Fig. 1. As mentioned earlier, speech noise reduction algorithms can be roughly divided into three basic categories based on the enhancement process. The rest of this section will explain typical noise reduction algorithms for these three categories.

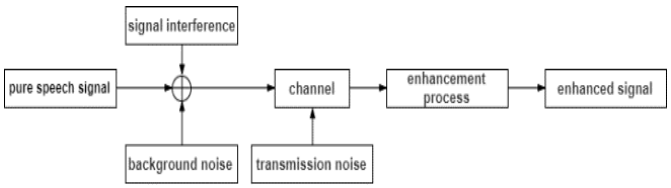


Fig. 1. The process of speech noise reduction algorithm.

### A. Filtering Technique: Spectral Subtraction

For the processing of additive noise, spectral subtraction is a widely used traditional and effective method [3]. The basic idea of the spectral subtraction is to subtract the noise power/frequency spectrum from the noisy speech power/ frequency spectrum to obtain a purified speech power/frequency spectrum. However, its premise is that the additive noise and the short smooth speech signal are independent of each other. Assuming that $s(t)$ is a pure speech signal, $n(t)$ is a pure noise signal, and $y(t)$ is a noisy speech signal, their signal sample values satisfy:

$$y(t) = s(t) + n(t) \tag{1}$$

Let $Y(\omega)$, $S(\omega)$, $N(\omega)$ represent the values after Fourier transform of the above signals, respectively, according to the Fourier transform equation:

$$DFT(x(k)) = \sum_{l=0}^{L-1} x(l) e^{\frac{-j2\pi kl}{L}} \tag{2}$$

Three signals in the frequency domain follows:

$$Y(\omega) = S(\omega) + N(\omega) \tag{3}$$

Spectral subtraction often uses the power spectrum. According to the assumption of spectral subtraction, the speech signal and the additive noise signal are independent and uncorrelated, so the power spectrum satisfies:

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 \tag{4}$$

We use $P_y(\omega)$ , $P_s(\omega)$ and $P_n(\omega)$ to represent the power spectrum of $y(t)$, $s(t)$ and $n(t)$, respectively. Then the (4) Can be rewritten as:

$$P_y(\omega) = P_s(\omega) + P_n(\omega) \tag{5}$$

The power spectrum of pure noise then can be estimated. Since the smooth noise power spectrum can be considered to be constant before the target sound, the so-called "silent period" before the sound to estimate the noise power spectrum, hence the estimated pure speech power spectrum is given by:

$$P_s(\omega) = P_y(\omega) - P_n(\omega) \tag{6}$$

The power spectrum obtained by this method can be considered as a relatively pure speech power spectrum, which can then be used to recover the speech time domain signal after noise reduction. At the same time, to avoid the negative power spectrum, $P_s(\omega)$ should equal to 0 if $P_y(\omega) < P_n(\omega)$ . The complete spectrum subtraction equation is as follows:

$$P_s(\omega) = \begin{cases} P_y(\omega) - P_n(\omega) & P_y(\omega) \geq P_n(\omega) \\ 0 & P_y(\omega) < P_n(\omega) \end{cases} \tag{7}$$

The entire spectral subtraction process is shown in Fig. 2.

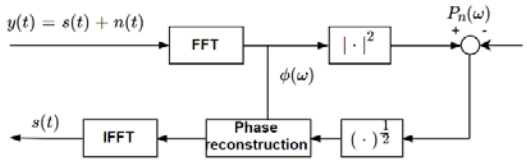


Fig. 2. The process of spectral subtraction.

There are many follow-up works based on spectral subtraction. Reference [4] aims to alleviate the problems of (7). The basic spectral subtraction uses 0 to set the negative value after spectral subtraction, which is also a reason why there is musical noise in the speech after spectral subtraction reconstruction. This work uses an over-subtraction technique for the noise spectrum and sets a lower limit for the value after spectral subtraction, instead of simply setting the negative value to zero. The improved (7) is as follows:

$$P_s(\omega) = \begin{cases} P_y(\omega) - \alpha P_n(\omega) & P_y(\omega) \geq P_n(\omega) \\ \beta P_n(\omega) & P_y(\omega) < P_n(\omega) \end{cases} \tag{8}$$

where $\alpha(\alpha \geq 0)$ is the over-subtraction factor and $\beta(0 < \beta < 1)$ is the lower limit parameter of the spectrum. In general, for high signal noise ratio (SNR), $\alpha$ should be a small value; for low SNR, $\alpha$ is a large value. Reference [5] proposed an MMSE-based spectral subtraction algorithm. It was pointed out that the over-subtraction factor and spectral lower limit parameters in (8) were determined through a large number of experiments and could not cover all noise cases, while the MMSE-based spectral subtraction algorithm can under the condition of the smallest square loss, the spectral subtraction parameters are optimally selected. In addition, some work has extended the scope of spectral subtraction from different perspectives [6-9].

### *B. Spectrum Reconstruction: MMSE-STSA*

MMSE-STSA (Minimum Mean Square Error Short Time Spectral Amplitude Estimator) [10] was proposed by Ephrain and Malah. We follow the notation in Section 2.1. Let $Y(k)$, $S(k)$, $N(k)$ denote the Fourier transformed spectrum of the noisy speech signal $y(t)$, the pure speech signal $s(t)$, and the noise signal $n(t)$. Among them, the amplitude of $Y(k)$ is $R(k)$, the phase is $\varphi_k$, the amplitude of $S(k)$ is $A(k)$, and $k$ is the frame index. Assume that the noise is stationary additive Gaussian white noise. According to the Bayesian formula, the estimated value $\hat{A}(k)$ of $A(k)$ is:

$$\begin{aligned} \hat{A}(k) &= \mathrm{E}\{A(k) \mid Y(k)\} \\ &= \int_0^\infty p[a_k \mid Y(k)] a_k da_k \\ &= \int_0^\infty \frac{p[a_k, Y(k)]}{p(Y(k))} a_k da_k \\ &= \frac{\int_0^{2\pi}\int_0^\infty a_k p[Y(k) \mid (a_k, \varphi_k)] d\varphi_k da_k}{\int_0^{2\pi}\int_0^\infty p[Y(k) \mid (a_k, \varphi_k)] d\varphi_k da_k}, \end{aligned} \tag{9}$$

where $\mathrm{E}(\cdot)$ denotes the expectation of the parameter; $p(\cdot)$ denotes the probability density function; $p(a_k)$ is the

probability density function of the amplitude $A(k)$ ; $p(a_k,\varphi_k)$ is the joint probability distribution of amplitude and phase.

According to the assumption of stationary additive Gaussian white noise:

$$p[Y(k)\,|\,(a_k,\varphi_k)]=\frac{\exp\left\{\frac{\left|Y(k)-a_k e^{j\varphi_k}\right|^2}{-p_n(k)}\right\}}{\pi p_n(k)} \tag{10}$$

$$p(a_k,\varphi_k)=\frac{a_k}{\pi p_s(k)}\exp\left\{-\frac{a_k^2}{p_s(k)}\right\},$$

where $p_s(k)$ and $p_n(k)$ denote the energy of the *k*-th spectral component of the pure speech signal and the noise signal, respectively:

$$\begin{aligned} p_n(k)&=\mathrm{E}\left\{\left|N(k)\right|^2\right\}\\ p_s(k)&=\mathrm{E}\left\{\left|S(k)\right|^2\right\}\end{aligned} \tag{11}$$

Hence, the estimated spectrum of pure speech $\hat{A}(k)$ can be obtained by:

$$\hat{A}(k)=\Gamma(1.5)\frac{\sqrt{v}}{\gamma_k}M(-0.5;1;-v_k)R(k), \tag{12}$$

where $\Gamma(\cdot)$ is the gamma function and $M(;;)$ is the confluent hypergeometric function. $v_k$ is defined as:

$$v_k=\frac{\xi_k}{1+\xi_k}\gamma_k, \tag{13}$$

where $\xi_k$ and $\gamma_k$ are the prior and posterior signal-to-noise ratios, respectively.

### *C. Model method: Regression Model*

Different from the previous two prior knowledge-based noise reduction methods, the model-based method directly models the noise reduction process and uses the model to establish the mapping relationship between noisy speech and pure target speech. In recent years, thanks to the development of deep learning, large-scale parameter learning has become possible, making model-based methods gradually become the mainstream in the field of speech noise reduction. Reference [11] proposed a method of spectral mapping, using deep neural networks to learn the mapping relationship between the noisy speech spectrum and the target pure speech spectrum. The framework of their model is shown in Fig. 3. This model is divided into three phases: a pre-training phase, a training phase, and an enhancement phase.

**Pre-training phase**. Considering that the DNN model is initialized with random initialization parameters, it will easily fall into a local optimal situation [12]. In the pre-training phase, the regression model uses a superimposed 3-layer Restricted

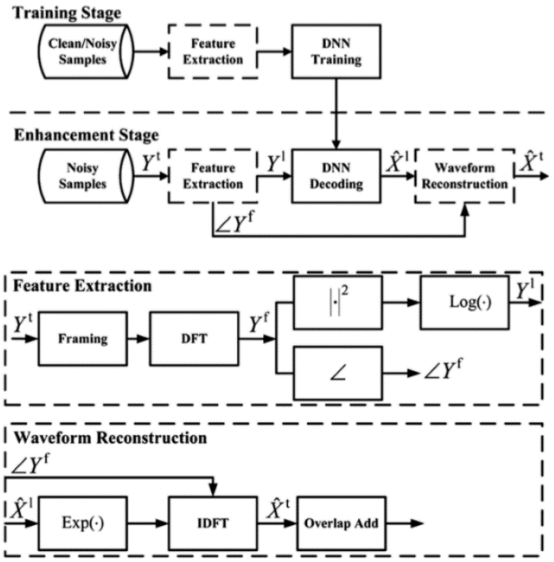


Fig. 3. Regression model framework

Boltzmann Machine to fit the noise data log spectrum [13]. The first layer is the visible layer of Gauss-Bernoulli RBM, that is, the data input layer, and the above two are the Bernoulli-Bernoulli RBM layers. The CD algorithm can effectively pre-train the model parameters.

**Training phase**. During the training phase, DNN-based regression models are trained using logarithmic power spectral features from pairs of noise and clean speech data. Logarithmic power spectral features are thought to provide perceptually relevant parameters, hence, first, short-time Fourier analysis is applied to the input signal, and the discrete Fourier transform (DFT) of each overlapping window frame is calculated. The logarithmic power spectrum is then calculated. This process is the same as the starting process of spectral subtraction. Finally, the stochastic gradient descent method (SGD) is used to optimize the mapping model under the mean square error function. Its loss function is as follows:

$$E=\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\left(\hat{X}_n^d\left(\mathbf{W}^l,\mathbf{b}^l\right)-X_n^d\right)^2, \tag{14}$$

where $\hat{X}_n^d\left(\mathbf{W}^l,\mathbf{b}^l\right)$ and $X_n^d$ represent the enhanced log spectrum and the original log spectrum of the *d*-th frequency component of the *n*-th speech frame, respectively; $\mathbf{W}^l$ , $\mathbf{b}^l$ represent the network weight and bias of the *l*-th layer. Assume that there are *L* network hidden layers, the *L*+1 layer is the output layer, and the training learning rate is $\lambda$ , then the network parameter update way is as follows:

$$(\mathbf{W}^l,\mathbf{b}^l)\leftarrow(\mathbf{W}^l,\mathbf{b}^l)-\frac{\lambda\nabla E}{\nabla((\mathbf{W}^l,\mathbf{b}^l))} \tag{15}$$

**Enhancement phase**. This phase uses the model to obtain the mapped speech spectrum, and reconstructs the speech signal by using the phase information of the extracted audio during the feature extraction phase and the overlap-add algorithm.

## IV. Performance Evaluation of Speech Noise Reduction Algorithm

The performance evaluation methods of speech noise reduction algorithms can generally be divided into two categories: subjective evaluation and objective evaluation. The subjective evaluation is to simply judge the denoised speech processed by the algorithm by randomly selected humans with normal hearing ability. The quality of this algorithm depends on the subject's preference for denoised speech. However, this evaluation method is time-consuming and the test is easily disturbed by environmental factors. Objective evaluation has the advantages of good stability and time-saving. A few common objective evaluation indicators are listed below.

### A. LSD

Log Spectral Distance (LSD) [14] is also called Log Spectral Distortion Measure, which is the distance measurement between two spectrums. The LSD between the spectrum $P(\omega)$ and the spectrum $\hat{P}(\omega)$ is calculated as follows:

$$LSD_{P(\omega)\sim\hat{P}(\omega)}=\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]} \tag{16}$$

In the evaluation stage, set $P(\omega)$ equal to the denoised speech spectrum and $\hat{P}(\omega)$ equal to the pure speech spectrum. The smaller the calculation result of the measure is, the closer the processed denoised speech is to the pure speech, and the performance of the algorithm is stronger.

### B. STOI

Short-time Objective Intelligibility (STOI) [15] is also a measure of the quality of denoised speech. The main idea of this method is to perform an STOI process on the pure speech signal and the restored signal respectively, and then calculate the correlation coefficient between the two. The magnitude of the correlation coefficient is positively related to the intelligibility of the speech signal.

Assume that *x* and *y* respectively represent the pure speech signal and the signal restored by the noise reduction algorithm. Then decomposed by the 1/3-fold band of the DFT [16]. The decomposition results of the pure speech signal are as follows:

$$X_j(m)=\sqrt{\sum_{k=k_1(j)}^{k_2(j)-1}\left|\hat{x}(k,m)\right|^2}, \tag{17}$$

where $k_1,k_2$ represent the 3x band boundaries, and $\hat{x}(k,m)$ represents the *k*-th DFT-bin of the *m*-th frame. Using the same process, the decomposition result $Y$ of the noise-reduced speech can be obtained. Normalize this signal to get $Y'$. The correlation coefficient between the two can be obtained by the following formula:

$$d_j(m)=\frac{\sum_n\left(X_j(n)-\frac{\sum_l X_j(l)}{N}\right)\left(Y_j'(n)-\frac{\sum_l Y_j'(n)}{N}\right)}{\sqrt{\sum_n\left(X_j(n)-\frac{\sum_l X_j(l)}{N}\right)^2\sum_n\left(Y_j'(n)-\frac{\sum_l Y_j'(n)}{N}\right)^2}} \tag{18}$$

where $d_j(m)$ denotes the correlation coefficient of the corresponding *m*-th frame. This value can be used to measure the intelligibility of the results obtained by the noise reduction algorithm.

### C. SegSNR

SegSNR is to calculate the signal-to-noise ratio for each frame of the noise reduction result and finally obtain the average signal-to-noise ratio. The larger the value of SegSNR, the better the noise reduction effect of the model. The SegSNR calculation process is as follows:

$$\text{SegSNR}=\frac{1}{M}\sum_{m=0}^{M}\lg\frac{\sum_{n=mN}^{mN+N-1}s^2(n)}{\sum_{n=mN}^{mN+N-1}\left(s(n)-\hat{s}(n)\right)^2} \tag{19}$$

where *M* is the number of frames and *N* is the frame length.

## V. Summary and Outlook

Speech noise reduction is an important branch of speech processing. Due to the different nature of noise, it is difficult to find a unified and robust method to solve this problem perfectly. Nonetheless, a lot of meaningful work has been proposed for different scenarios. This article first summarizes these typical methods from the perspective of three categories and then describes several more common noise reduction algorithm evaluation methods.

For the future research direction of noise reduction, we believe that: due to the great success of deep learning in various fields, the performance of the model has surpassed the traditional machine learning methods. Second, traditional algorithms (such as spectral subtraction) are not suitable for a large number of Data learning, so the research focus of the whole noise reduction algorithm will be biased towards deep learning.

Furthermore, the current deep learning-based noise reduction models, such as the DNN model mentioned in section 2.3 of this article, as well as the CNN model and sequence-based LSTM model not mentioned in this article, are almost always based on learning a slave noise the mapping function of speech spectrum to pure speech spectrum. This learning method lacks a priori information. By analyzing the sensory knowledge of the human ear's noise and human voice, we know that a priori knowledge is important for signal

separation and enhancement (In terms of biological hearing, this a priori is to determine which target signal is from the mixed-signal), so we consider that combining voiceprint recognition and noise reduction is a direction that can be explored and studied.

## ACKNOWLEDGMENT

This work was supported by Guangzhou Power Supply Co., Ltd (Major special projects of China Southern Power Grid), under Grant No. GZHKJXM20170059.

## REFERENCES

[1] J. Liu and X. Xiang. "Review of the anti-noise method in the speech recognition technology." 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2017.

[2] J. Benesty, M. M. Sondhi, and Y. Huang. "Springer handbook of speech processing," Springer, 2007.

[3] 3. S. Boll. "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, 27(2):113–120, 1979.

[4] M. Berouti, R. Schwartz, and J. Makhoul. "Enhancement of speech corrupted by acoustic noise," IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 4, pages 208–211. IEEE, 1979.

[5] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan. "A parametric formulation of the generalized spectral subtraction method," IEEE transactions on speech and audio processing, 6(4):328–337, 1998.

[6] P. Sovka. "Extended spectral subtraction-description and preliminary results," Research report, 1995.

[7] H. Gustafsson, S. E. Nordholm, and I. Claesson. "Spectral subtraction using reduced delay convolution and adaptive averaging," IEEE Transactions on Speech and Audio Processing, 9(8):799–807, 2001.

[8] C. He and G. Zweig. "Adaptive two-band spectral subtraction with multi-window spectral estimation," IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. volume 2, pages 793–796. IEEE, 1999.

[9] N. Upadhyay and A. Karmakar. "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," Procedia Computer Science, 54:574–584, 2015

[10] Y. Ephraim and D. Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee. "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1):7–19, 2015..

[12] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks," Science, 313(5786):504–507, 2006.

[13] Y. Bengio. "Learning deep architectures for ai," Foundations and trends® in Machine Learning, 2(1):1– 127, 2009.

[14] L. R. Rabiner, B.-H. Juang, and J. C. Rutledge. "Fundamentals of speech recognition," Pearson Education India, 2008.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "A short-time objective intelligibility measure for time-frequency weighted noisy speech," IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4214–4217. IEEE, 2010.

[16] N. Li and P. C. Loizou. "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," The Journal of the Acoustical Society of America, 123(3):1673–1682, 2008