A LITERATURE SURVEY ON SINGLE CHANNEL SPEECH ENHANCEMENT TECHNIQUES

Naik D C, A Sreenivasa Murthy, Ramesh Nuthakki

Abstract: Speech enhancement deals with the handling of noisy speech signals in order to improve people's perception or better system understanding when noise destroys speech information. It is usually difficult to keep speech undistorted while reducing noise and thus limiting the performance of speech enhancement systems— the compromise between distortion of speech and reduction of noise. With noisy speech with medium to high SNR, the goal will be to generate subjectively realistic signal by reducing noise levels, and for those with low SNR, the goal could be to reduced noise level while retaining intelligibility. In this work, discussion on the need for speech enhancement, its applications, and an overview of classification and various approaches available has also been given and done an extensive literature survey on speech enhancement techniques with different platforms.

Index Terms— Single-channel speech enhancement; spectral subtraction; deep neural networks; performance measures

1 INTRODUCTION

Speech can be thought of as the most common and desirable medium for humans to communicate with each other. In addition to the inter-human communication, speech has found a lot of applications in the human-machine interface, thanks to the advancements in technology over the decades. For any communication medium, it is always desirable that noise has little or no effect on it and that too for a medium as popular as speech, it must retain unaffected by noise. But, keeping an eye on the various kinds of noises and their sources, it is difficult to acquire a speech that is free from noise. So, the problem is clear that the effect of noise has to be made zero or the minimum possible given a speech available has been noisy, degraded by noise. The research in this aspect is popularly known as speech enhancement. Alternatively, speech enhancement can be defined as a process employed for improving perceptual aspects of speech like quality, intelligibility, or degree of listener fatigue. Many researchers over the decades have come up with many speech enhancement algorithms intended in mitigating the effect of the noise on speech. However, the complexities of speech signal mean that this research still presents a considerable challenge. There is even a trade-off between speech distortion and noise reduction which limits the performance of speech enhancement systems. That is, it is usually challenging to reduce noise without distorting speech.Applications of speech can be broadly classified into two categories, one is the inter-human communication and the other being human-machine interface. The former can be considered as two people transferring messages in acoustic form in two cases, one where both are near to each other in a noise-free environment and it can be observed that communication is generally easy and accurate, and another case where they are at a distance from each other with noisy background, the listeners ability to understand suffers. In a human-machine interface, speech is usually recorded, or processed in electrical form; the conversion media, as well as transmission media, generally introduce distortion, resulting in a noisy speech signal. This distortion reduces the intelligibility or quality or both of a speech. So, it often becomes necessary to process noisy speech for noise removal using a speech enhancement system. As the performance of speech coders and voice recognition gets degraded due to the presence of noise incorporation of speech enhancement in these systems is common. The objectives of speech enhancement is to reduce the noise level, increase intelligibility and reduce auditory fatigue.

The speech enhancement techniques can broadly be classified as a single channel, dual channel and multi-channel enhancement techniques depending on how many acquisition channels are used. If the signal acquisition is made using a single microphone and an enhancement technique is devised to enhance the signal, such an enhancement technique is called a single-channel speech enhancement technique. In cases where two acquisition channels are employed, one for recording speech and the other to get reference signal for noise, enabling us to use adaptive noise cancellation, the enhancement technique is called dual-channel enhancement technique. Microphone arrays are utilized in multi-channel enhancement techniques to make phase alignment for rejecting the undesired noise components. Single-channel speech enhancement techniques are the most challenging enhancement techniques as there is no reference signal for noise available. One more classification of speech enhancement techniques can be given based on the type of processing they go through. In spectral subtraction methods, an estimated noise magnitude/power is subtracted off from the noise speech magnitude /power respectively while the phase remains undisturbed. In Wiener filtering methods, a spectral gain function of SNR measures is computed and applied to the noisy speech to get an estimate of clean speech. Statistical model-based methods utilize asymptotic statistical properties of the Fourier expansion coefficients to derive a spectral amplitude estimator. Signal subspace methods essentially represent the application of a principal component analysis (PCA) approach to ensembles of observed time-series obtained by sampling.The spectral subtraction is a well-known

- *Naik D C is currently working as a Full time Research Scholar in the department of Electronics and Communication Engineering, UVCE, Bangalore University, Bengaluru, India, 9611445531. E-mail: chethan.naik24@gmail.com*
- *A Sreenivasa Murthy currently working as a professor in the Department of Electronics and Communication Engineering, UVCE, Bangalore University, Bengaluru, India, 9900682076. E-mail: uvceasm@gmail.com*
- *Ramesh Nuthakki currently working as an Asst. Professor in Atria Institute of Technology, Bengaluru, India, 9448476272. E-mail: nuthakki.ramesh@gmail.com*

method to reduce the background noise from noisy speech signal which assumes noise & signal are additive and are uncorrelated. Hence an estimate of noise obtained from the noise only region of noisy speech is subtracted from the noisy speech to obtain an estimate of the speech signal in which speech quality & intelligibility are enhanced. However, in the normal speech communication scenario noise is mostly colored and does not impact the entire speech spectrum uniformly.Single-channel speech enhancement plays an important role in communication systems, particularly in a noisy environment, as this technique is popularly used due to an advantage in computing that easily and significantly reduces background noise. Because single-channel speech enhancement algorithms based on statistical method assume that the noise signal in the time domain is statistically stationary.Recent studies focused on a non-linear approach to the process of subtraction. Thanks to the fluctuation in signal-to-noise ratio across the speech spectrum, this strategy is justified the distribution of real-world noise is not linear, unlike white Gaussian noise with a flat spectrum. Therefore, the noise signal has no consistent impact on the entire spectrum of the speech signal. Many frequencies are more adversely affected than others. Of example, the low frequencies, where most of the speech power resides are more influenced than the high frequencies in multi-talker babble. Therefore, it becomes necessary to estimate an appropriate factor that will only eliminate the required amount of noise spectrum from each frequency bin (ideally) to avoid disruptive speech subtraction while eliminating most residual noise.This work reflects on improving speech signals degraded by statistically unbiased additive noise. For speech enhancement, the output of only one microphone containing the noisy signal is assumed. Speech enhancement aims to reduce the level of noise in a distorted speech signal while reducing the addition of sound defects to improve the quality and/or intelligibility of speech. Most solutions to speech enhancement are based on the framework of Analysis-Modification- Synthesis (AMS), where a short-term Fourier transform (STFT) is used in the part of the analysis to improve performance in the spectral domain. Discrete-time noisy signal is segmented into short-term frames using the 50% overlap Hanning window. The earlier noisy speech stage is applied to the changed magnitude / power domain and the time waveform is synthesized from it. To get the enhanced speech the synthesized short-time waveforms are applied with a 50% overlap between adjacent frames.

2 LITERATURE SURVEY

S F Boll [1] proposed a technique for reducing background noise is the spectral subtraction algorithm and is a very old basic effective method for reducing background noise from the noisy speech signal. When designing this algorithm, some of the hypotheses are considered, i.e. the background noise is added to speech acoustically or electronically. The background noise condition stays constant locally to the degree that its expected value of spectral magnitude is equal to its expected value after speech activity just before speech activity. Through subtracting the noise amplitude scale from the noisy speech spectrum the technique was used to approximate the magnitude frequency spectrum of the corresponding clean speech The measurement of noise amplitude range is calculated by combining the first few frames of non-speech activity from the noisy speech signal and the secondary procedures applied to reduce the effect of background noise by half-wave rectification, residual noise reduction, and additional signal attenuation during non-speech activity. The diagram of the framework is as shown in Figure 1.

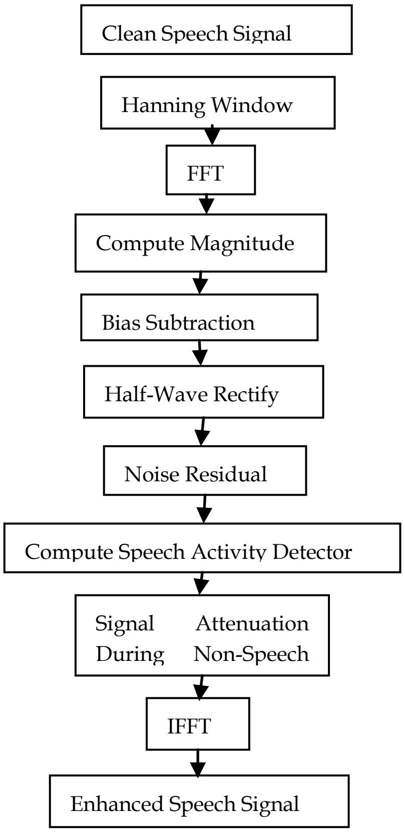


Fig. 1. Block Diagram of Spectral Subtraction.

The result of half-wave rectification is that the noise distortion calculated at that frequency would filter down the magnitude scale at each frequency. The bias value can be modified from frequency to frequency and from time window to window of study. The noise residual's noticeable effects can be reduced by using its frame-to-frame randomness. In fact, at a certain frequency bin. Since the noise residual at each frame of the analysis will randomly fluctuate in amplitude, it can be removed by replacing its current value with the minimum value selected from the neighboring frames of the analysis. Signal control during non-speech activity is the last change in noise reduction as the equilibrium must be established between the frequency and characteristics of the noise perceived during speech activity and the noise sensed during speech absence.

M. Beroutiet.al[2] suggested a strategy for optimizing speech distorted by broadband interference by taking into account two considerations in its implementation: first, the Author excludes an α factor (over-subtraction factor) from the noise spectrum where α 1 differs from frame to frame. Second, we prohibit the filtered signal's spectral components from going below a certain minimum β (spectral floor) pre-set level. The method involves subtracting from the speech power spectrum an estimate of the noise power spectrum, setting negative discrepancies to zero, recombining the new power spectrum with the initial phase, and then reconstructing the time waveform. While this approach eliminates broadband interference, distracting "musical distortion" is usually added as well. The researcher suggested this approach to remove this "musical noise" and is as shown in Figure 2.

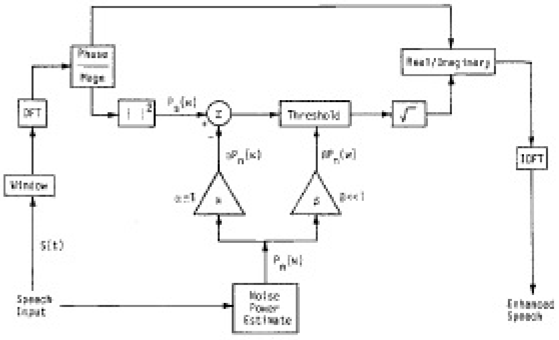


Fig. 2. Block Diagram of power Spectral Subtraction Domain

Yariv Ephraim and David Malah [3] suggested a voice amplification algorithm corrupted by irrelevant additive noise when there is only the distorted speech signal. The basic approach is to optimally approximate (under the criteria of MMSE (Minimum Mean Square Error) and an expected mathematical method, the short-term spectral amplitude (STSA) and the complex exponential of the speech signal process. Author uses this method to measure the two elements of the short-term Fourier transformation (STFT) independently as optimally as possible rather than estimating the STFT itself optimally. The researcher has shown that the STSA and the exponential complex cannot be optimally calculated simultaneously. The researcher then uses an optimized MMSE STSA estimator and blends it with an optimal MMSE estimator of the complex phase exponential that does not impact the calculation of STSA. The researcher found that at high SNR, the estimator of MMSE STSA and the estimator of Wiener STSA resulting from the optimum estimator of MMSE STFT are almost equal.The aim of this paper [4] is to incorporate recent research on model-based speech enhancement and provide a coherent quantitative context for specific speech enhancement issues. The composite reference model is the most common statistical model of speech signals and has proved to be extremely useful in applications of speech recognition and enhancement. Such arrangement is one of the most difficult situations in speech enhancement as no reference signal is believed to be available to the noise and the clean speech cannot be pre-processed before the noise is affected.In the suppression law suggested by Ephraim and Malah, Olivier Cappe [5] offered an overview of the various mechanisms that predict the effect of musical noise. The key element was found to be the nonlinear smoothing method used to produce a more accurate SNR approximation. The use of the smoothing technique does not produce audible noise in the signal with a sufficient parameter range. Nevertheless, during abrupt transients, low-level signal components experience a measurable over-attenuation. This transient distortion is hardly appropriate, and more detailed hearing experiments would be needed to determine whether or not it would be beneficial to use an overlap factor greater than 50%. Eventually, the attenuation function suggested by Ephraim and Malah has been shown to avoid the appearance of the musical noise effect even when the background noise is improperly stationary.Pascal Scalarti and Jozue Vieira Filho [6] show a unified overview of the frequency domain's primary single microphone noise reduction technique. The researcher has proposed a new approach to use the A Priori Signal to Noise Ratio to improve speech. The new solution allows a substantial reduction in noise power without the risk of ‘musical-noise’.S. D Kamath and P Loizou [7] suggested a multi-band spectral subtraction method that takes different frequencies into account the idea that colorful noise affects the speech spectrum. Based on the frequencies the voice spectrum is separated into N non-overlapping bands, and spectral subtraction is done in each band separately.Cyril Plapous.et.al [8] suggested a methodology for noise reduction based on a priori SNR estimation in two stages. In the first step, the a priori SNR approximate gives interesting properties, but suffers from one frame delay that is eliminated by the second step i.e. TSNR algorithm (Two-step Noise Reduction). The researcher proposes to estimate the multiplicative gain G (p;k) in a two-step method in order to improve the efficiency of the noise reduction system. This method is referred to as the algorithm (TSNR). In the first step, we calculate the SNRpriodd (p;k) and/or SNRinst (p;k) multiplicative gain function. This method is referred to as the algorithm of decision-direct (DD). The multiplicative gain in the first step is then used to optimize the estimation of the a priori SNR. Therefore, this technique will monitor the non-stationarity of the speech signal instantly without adding the musical noise effect that is demonstrated in the speech communication linguistic sense.Yi Hu and Philipos C. Loizou [9] provided a measure of performance to evaluate the algorithms implemented to determine which algorithm performs well in terms of subjective and objective measures. The measure of subjective performance is called a listening test. Which were designed by Dynastat, Inc. according to ITU-T recommendation P.835. In a subjective test, the P.835 methodology was designed to reduce the uncertainty of the listener as to which component(s) of a noisy speech signal, i.e., speech signal, background noise, or both, should form the basis of their overall quality ratings. This method provides the listener with instructions to listen to speech signals and rate the enhanced speech signal on 1. Speech alone with a 5-point signal distortion index (SIG) (Table 1), 2. Background noise on its own using a natural intrusiveness (BAK) level of five levels (Table 2), 3. The overall effect using the OVRL scale-[1=bad, 2=poor, 3=fair, 4=good, 5=excellent].

Table I: Scale of Signal Distortion (SIG)

5	Very Natural, no degradation
4	Quite natural, slight degradation
3	Slightly natural slightly degraded
2	Quite unnatural, quite degraded
1	Very natural, degraded

Table II: Scale of Background Intrusiveness (BAK)

5	Not visible
4	Slightly visible
3	Noticeable but not intrusive
2	Quite noticeable, a little intrusive
1	Really interesting, really intrusive

The method of rating the enhanced speech signal and background of the enhanced speech signal has been designed to lead the listener to combine the results of both the clean speech signal and background noise and render their overall quality scores. Listeners successively used one of the three five-point score scales (SIG, BAK, and OVRL) to report their decisions on the value of the test environment for each experiment within the triad. In addition to the experimental conditions, each experiment included several comparison criteria intended to differentiate the SIG, BAK, and OVRL scores of the listener individually over the entire rating scale five-point spectrum. It is more specifically used to evaluate the intelligibility of the speech signals.Objective performance measures are usually determined using some mathematical formula from the original undistorted voice signal and the improved voice signal. It doesn't require people to listen, so it's less expensive and less time-consuming. Objective tests are often used to achieve an approximate value estimate. Then these measurements are used iteratively to check arbitrary test conditions for reliability. Some successful estimators have been established for objective measures, but at some stage we still need to assess qualitative value as there are still cases where approximations fail. Many objective measures are highly correlated with perceived subjective consistency, while others are more correlated with subjective intelligibility. Some of the objective measures include the following 1) Standard Segmental SNR measures (dSEGSNR)2) Itakura-saito Distortion Index (dIS) 3) Log-likelihood ratio (dLLR) 4) Log-area ratio(dLAR) 5) Weighted time and frequency domain SNR measures 6) Perceptual evaluation of speech quality(PESQ) 7) Weighted spectral slope(dWSS).Cyril Plapous et.al [10] proposed a decision-direct (DD) method that dramatically limits the level of musical noise, but the approximate a priori SNR is biased as it depends on the estimation of the speech spectrum in the previous frame. The gain function therefore matches the previous frame rather than the current one that degrades the performance of noise reduction. The consequence of this discrimination is an irritating effect of reverberation. The author suggests a method called a two-step noise reduction (TSNR) technique that will solve this problem while maintaining the advantages of the decision-direct approach. A second step to remove the bias of the DD approach, thereby removing the reverberation effect, refines the approximation of the a priori SNR. Nonetheless, traditional short-term noise reduction methods, like TSNR, apply harmonic distortion of improved speech signals due to small signal-to-noise ratio estimators ' unreliability. This is mainly due to the hard work of approximating the noise power spectrum density (PSD) in the single microphone scheme. The author suggests a method called Harmonic Regeneration Noise Reduction (HRNR) to overcome this problem. An approach to nonlinearity is used to effectively regenerate the degraded harmonics of the distorted signal.For mostly non-stationary noise conditions, S Rangachari and philipos C Loizou [11] suggested a noise estimating algorithm. The noise approximation is modified by using time and frequency-dependent smoothing factors to average the noisy speech power spectrum, which are aligned in individual frequency bins based on the probability of signal-presence. Signal intensity is calculated by measuring the ratio of the noisy sound power spectrum to its local average, which is continuously updated with a look-ahead parameter by comparing previous values of the noisy speech power spectrum. The local minimum algorithm for approximation adapts to extremely non-stationary noise conditions very easily.Ningpin Fan Justinian.et.al [12] suggested an Enhanced MCRA (minimum controlled recursive average) algorithm (EMCRA), which indicates less voice leakage and quicker response time to match abrupt changes in the range of noise energy. The MCRA approach [35] indirectly uses spectral minima. This approximates noise by combining past spectral power values with a factor of smoothing that is modified by the likelihood of signal occurrence in sub-bands. This, in effect, is determined by the ratio within a specified time period between the noisy speech's local energy and its spectral minima. Up to a certain threshold, the ratio shows the absence of speech with a lesser value. Temporal smoothing is also done to reduce fluctuations between speech and non-speech segments, allowing use of a strong correlation with speech activity in adjacent frames. The MCRA method is computationally efficient, robust with respect to the input of the SNR and the type of additive noise underlying it.Yi Hu and Philipos C. Loizou [13] show how, in terms of predicting the quality of noisy speech signal enhanced by noise suppression algorithms, to evaluate the performance of several objective measures. Objective performance measurements are calculated by the mathematical formula which is considered to be a wide range of distortions introduced by different types of real-world noise at different signal-to-noise ratio levels by four classes of speech enhancement algorithms: spectral subtractive, subspace, statistical-model-based, and Wiener algorithms. The qualitative quality scores are achieved using the technique ITU-T P.835 developed to assess the quality of the improved speech signal together with the three-dimensional clean speech sound: signal distortion, noise distortion, and overall quality. Many new hybrid objective measures are also introduced using nonparametric and parametric regression analysis techniques to incorporate the individual objective measures.This paper [14] includes several classical and modern spectral approximation algorithms that are integrated into the formal spectral subtraction methods to improve speech enhancement efficiency. This paper suggests three methods: the SSW method (combined method based on spectral subtraction and weld estimator), the SSY approach (combined method based on YULE-walker AR estimator) and the SSM method (combined method based on MUSIC (multiple signal classification) estimators).The Welch method is an improved periodogram method, which is one of the most effective methods for classical spectral approximation. One of the parametric methods for conventional spectral subtraction calculation is the Yule-Walker AR (auto-regressive) technique. This spectral method measures the parameters of the AR by constructing a biased estimate of the autocorrelation function of the signal and solving the minimization of the forward prediction error by the least squares. One of the non-parametric methods for conventional spectral analysis is the MUSIC (multiple signal classification) technique. It is techniques of frequency estimators based on the autocorrelation matrix's Eigen analysis.Yang Lu and Philipos C Loizou [15] proposed a new approach-based (GA) method for spectral subtraction. It addresses the two major weaknesses of spectral subtraction mentioned above: musical noise and invalid assumptions about zero cross-terms. The method used is largely deterministic and is based on representing the noisy speech spectrum as the sum of the clean signal and noise vectors in the complex plane. It can provide valuable insights into the spectral subtraction method by geometrically reflecting the noisy continuum in the complex plane. For one, such a geometric point of view can provide upper limits on the disparity between the noisy and clean spectra stages. It will also inform us whether, considering the distorted speech spectrum it is potentially possible to recover exactly the clean signal magnitude.Jomg Mo Kum et.al [16] proposed a method based on a conditional maximum a posteriori (MAP) criterion to improve the performance of minima controlled recursive averaging (MCRA). From an investigation of the MCRA scheme, it is discovered that the MCRA method cannot take full account of the inter-frame correlation of voice activity since the estimate of noise power is adjusted by the probability of speech presence depending on the current frame observation. In order to avoid this effect, the proposed MCRA approach uses the conditional MAP test in which the calculation of noise intensity is derived using the likelihood of speech existence based on both the present measurement and the decision of speech activity in the previous frame.Kuldip Paliwal et.al [17] shows that modulation domain and acoustic domain model is suitable and effective for spectral subtraction. For this purpose, the author extends to include modulation domain processing the traditional modification-synthesis framework of the analysis. By applying the spectral subtraction algorithm in the modulation domain, the author then remunerates the noisy modulation spectrum for additive noise distortion. Typically associated with acoustic spectral subtraction, the proposed modulation spectral subtraction does not suffer from the musical noise artifact. The author looks at a fusion of modulation spectral subtraction with the MMSE method to achieve further improvements in speech quality. In the short-time spectral field, the fusion is done by integrating the magnitude range of the above algorithms for speech enhancement.Typical modulation filtering strategies are concerned with two key constraints. Next, they use a filter design based on the language modulation spectrum's long-term properties, thus neglecting the noise properties. As a result, the noise components found within the regions of voice modulation are not excluded. First, the modulation filter is constant and added to the whole signal, while over time the features of speech and noise change.Hou Xuchu and Zhu Xiaojing [18] address the phenomenon of harmonic distortion in traditional methods of short-term noise suppression, especially in noisy environments with poor signal-to-noise ratio (SNR). To solve this problem, the researcher suggested a basic but efficient method of harmonic regeneration. Using traditional methods of short-time noise suppression, the complete harmonics of speech are achieved by applying a non-linear function to the improved expression. The artificial signal is then rectified by updating the amplitude of the spectrum in three different sub-bands of the frequency.

Jesper Rindom et.al [19] has suggested a template that is intended specifically for voiced speech signal and is not appropriate for voiceless speech signal. That is, for some portions of the speech signal-dependent methods based on the signal statistics can add unwanted distortion compared to signal-independent methods based on the noise statistics. Because both signal-independent and signal-dependent solutions to voice enhancement have their advantages and disadvantages, integrating them is important to reduce the impact of their disadvantages. It supports the shared use of such sorting processes, which from a practical point of view can be helpful. The researcher states that both signal-independent and signal-dependent methods have benefits and are closely related by the experimental results. In fact, it is very useful as part of experiments by essentially integrating signal-independent and signal-dependent enhancement methods when adding both approaches to the real-life speech signal together.When discussed in the above section, there are additional advantages and disadvantages in the signal-independent and signal-dependent filter design approaches. Hence, exploring whether these methods can be merged to achieve the benefits of both while reducing the impact of their drawbacks is highly relevant. They provide further insight into the relationship between the signal-independent and signal-dependent filter design methods as a first step in this direction. More precisely, the researcher considers the relationship between two recently proposed filter models, namely the orthogonal decomposition-based minimum distortionless response (ODMVDR) filter [36] and the linearly restricted minimal variance (HDLCMV) filter of harmonic decomposition [37]. The ODMVDR filter is independent of the signal, while the HDLCMV filter is dependent on the signal. In addition, when the target signal is intermittent, we provide some closed-form output measurements for filters constructed using both signal-independent and signal-dependent approaches.Navneet Upadhyay and Abhijit Karmakar [20] proposed an improved multi-band spectral subtraction algorithm to improve speech quality in different noise environments such as babble noise, car noise, helicopter noise and random noise. The suggested algorithm splits the whole speech spectrum into different uniformly distributed continuous frequency bands and independently performs spectral over-subtraction in each band. Without using speech delay detection, the proposed algorithm uses a novel approach to continuously estimate the noise from each band. The noise in each uniformly spaced frequency band is approximated and modified by adaptively smoothing the noisy signal power. A linear function of the a-posteriori signal-to-noise ratio (SNR) controls the smoothing parameter. Instead, using the appropriate over-subtraction variable in each band, the spectral over-subtraction is used to eliminate background noise. The suggested algorithm incorporates further noise reduction so that the residual noise does not disturb the listener while reducing the distortion of expression added during the process of enhancement. This proposed algorithm is well adapted for the form of speech-shaped noise, as shown in Figure 3

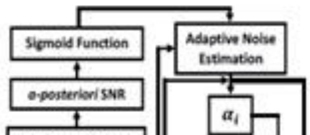


Fig. 3. Block Diagram of Multi-Band Spectral Subtraction.

By applying a harmonic model to the enhancement algorithm, Martin Krawezyk and Timo Gerkmann [21] discuss a method for reconstructing the spectral phase of the voiced speech signal. Where the baseband phase gap shows structures present in the clean speech spectral process and is reconstructed using the proposed algorithm. The basic concepts were pointed out, as well as the importance of improving the spectral phase, and the researcher showed that noise between harmonics of voiced speech can be effectively suppressed only by reconstructing the spectral phase. In addition to the sole enhancement of the spectral phases presented in [38], the researcher demonstrated that the suggested reconstruction process could also be paired with spectral amplitude estimators to further improve speech efficiency by increasing background noise. In addition, the restored phase provides valuable data that can be used for enhanced phase-sensitive amplitude estimators or even complex spectral coefficients estimators. These combinations may outperform traditional enhancement schemes based on amplitude. This describes the drawback of phase-based noise reduction.Josef Kulmer and Pejman Mowlaee [22] suggested a phase estimation approach based on the degradation of the instantaneous noisy phase spectrum into its underlying components: minimum phase (vocal tract filter), linear phase (fundamental frequency) and dispersion phase. Upon subtracting the linear phase from the instantaneous noisy phase, the author shows that a temporal smoothing filter can be added to the remaining unwrapped phase without using a predefined limit to reduce the noise contribution. The Author tested the efficacy of the suggested phase estimation approach in two scenarios: explicitly on the noisy speech and in accordance with an amplitude enhancement scheme as a post-processor. Consistent improvement is achieved in perceived performance as well as in intelligibility.The multi-stream (MS) method has been wildly used in automatic speech recognition (ASR) for many years and has been shown to be effective ways to enhance the accuracy and robustness of recognition ASR systems. Based on the fact that background noises or discrepancies do not affect the multiple data streams equally, the MS system usually outperforms single streams in diverse and unpredictable noisy environments by accurately identifying and fusing complementary data streams. The main difficulty in using the multi-stream approach to speech enhancement is that it is not possible to extract the speech waveform directly from many forms of data streams, such as visual lip information.Yan Xiong.et.al [23] proposed a new model for multi-stream speech enhancement that is used by multi-stream information even when some of the data streams are not directly concerned with speech waveform. The approach proposed is based on the single-channel model-based speech enhancement technique, with the exception of using a multi-stream system to distinguish the noisy speech frame. Based on the results of identification, a class-dependent filter improves the noisy speech as the traditional model-based methods of enhancement. In this way, multi-stream information does not need to be used to directly retrieve the voice waveform, but to improve the frame classier's robustness, which can outperform conventional model-based enhancement methods by extracting single data stream from noisy acoustic voice signals.Betty Kurian.et.al [24] suggested a PNCC (Power Normalized Cepstral Coefficients) which is a methodology for removing characteristics using nonlinearity of power-law. Time analysis is used in the PNCC processing method to estimate the degradation of the environment. It uses nonlinearity of power-law which approximates the non-linear relation between the frequency of the signal and the rate of auditory nerve firing. It uses temporary masking to suppress noise and to support real-time online processing. Speech signal has a high spectrum of modulation and attempting to speak power varies rapidly from components of noise. Thus, the parameters characterizing environmental degradation are analyzed using a longer window of 50-120ms duration. The processing of PNCC has three levels of preliminary processing, environment compensation and final processing. Initial processing is similar to conventional processing of MFCC except that the frequency analysis is carried out using a bank of gamma tone filters. Gamma tone filters are used to design auditory filters and gamma tone weighting improves accuracy. Speech enhancement is done in the second stage. Using longer-term temporal analysis, nonlinear time-varying operations are conducted to improve robustness.Qiquan Zhang and Mingjiang [25] suggested a robust speech enhancement algorithm for non-stationary noise conditions, consisting of a multi-band spectral subtraction estimator (MBSS) and a minima controlled recursive average (MCRA) noise approximation as shown in Figure 4.

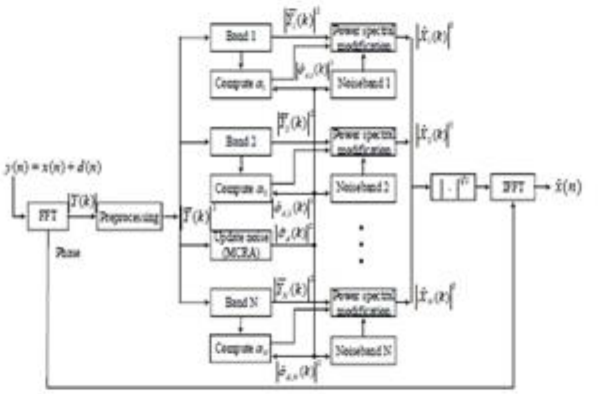
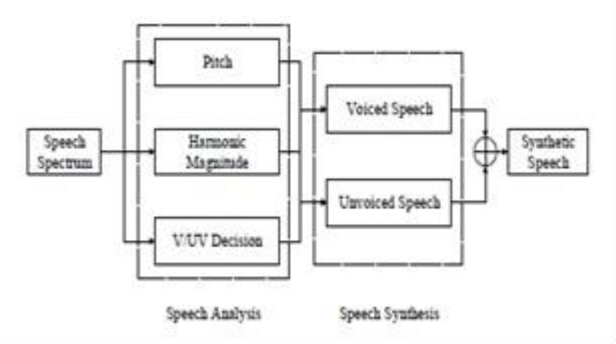


Fig. 4. Block Diagram of the MBSS and MCRA design.

A multi-band method is implemented by dividing the whole spectrum into multiple sub bands (about 3 bands) and independently applying spectral subtraction in each band. The approximate noise takes advantage of the observation that the noise signal typically has a non-uniform effect on the speech spectrum and is given by averaging the past power spectrum value and using time and frequency-dependent smoothing factor calculated based on the probability of speech presence in sub bands [39]. The author uses a minima controlled recursive averaging (MCRA) method in this paper to estimate the noise spectrum. The algorithm for noise estimation incorporates the simplicity of the recursive averaging process with low monitoring robustness. MCRA algorithm can detect sudden changes in the spectrum of noise easily and reduce the complexity of the computation.Zhang Wenlu and Peng Hua [26] suggested a revised Wiener speech enhancement algorithm with phase spectrum correction to improve the performance of distinctive Wiener speech enhancement algorithms in low SNR. Since many speech enhancement algorithms have always used the assumption that the measured noisy speech phase range remained unchanged specifically as an improved speech phase spectrum and estimated noise frequency spectrum due to the normal noise power spectrum of speech delays, rendering it inaccurate for eventual signal processing. To solve the above problems, the researcher suggested a method that updates the Wiener filtering estimate of noise energy spectrum based on whether the current frame to be calculated is speech signal or non-speech signal and combines the amplitude estimation of clean speech signal with phase spectrum compensation which preserves more speech signal data.The updated noise spectrum estimation will be determined as follows. Next, the voice start frame is calculated by identifying noisy speech sound activity and determining the initial value of the noise power spectrum by averaging the noise power over the original non-speech activity frames. The next step is to analyze the current frame to confirm whether it is a speech signal or non-speech signal and to change the noise energy spectrum estimate. If it is a non-speech signal, the researcher measures the current frame's noise power spectrum by smoothing the noise power spectrum from the current frame's initial noise signal power spectrum values.Nahma L et.al [27] suggested an adaptive smoothing factor for the updated a priori SNR estimation method in which the smoothing factor is regulated by a time-frequency weighting variable with a sigmoid form. Flexible smoothing to the a priori approximation of the SNR is therefore calculated to boost the monitoring system for sudden changes in the instantaneous signal to noise ratio. The author discovered that while reducing musical noise, the adaptive weighting factor helps preserve the weak speech components. The suggested solution improves the DD a priori SNR estimator by introducing an adaptive smoothing parameter dependent on the instantaneous a posteriori SNR and an adaptive sigmoid feature to improve speech onset monitoring. The approach reveals by the experimental results that weak speech components can be maintained relative to two other comparison approaches discussed in this paper and by preserving the DD method's benefit in minimizing musical noise. The benefit of this approach is that a priori SNR estimation will reduce a one-frame delay by considering a smoothing factor nearer to 1.

Siddala Vihari et.al [28] performed a comparative analysis of the approach to spectral subtraction and Wiener filtering when non-stationary noise renounces expression. All algorithms are analyzed by the researcher using stationary noise scenario. Decision-directed (DD) approach is used to estimate the time-varying noise spectrum resulting in improved intelligibility performance and reduced musical noise. However, the current frame's a priori SNR estimator depends on the preceding frame's estimated speech spectrum. The undesirable implication is that the gain feature does not suit the current frame, resulting in a distracting echoing effect causing a bias. A procedure called a two-step noise reduction (TSNR) algorithm has been applied to solve the problem that immediately measures the signal's non-stationarity but does not lose the DD approach's benefit. An additional step to eliminate the bias has altered and enhanced the a priori SNR estimate removing the reverberation effect. Even with TSNR, the output obtained still suffers from harmonic distortions inherent in all techniques of short-term noise suppression, the main reason being the inaccuracy in estimating PSD in single-channel systems.An improved version of the fusion strategy was introduced by Julien Bosco and Eric Plourde [29] using the MMSE SMM (Minimum Mean Square Error Short-Time Spectral Modulation Magnitude) estimator to boost the spectral modulation domain instead of using the SMS (Spectral Modulation Subtractor) estimator, the former being aggregated with the MMSE SA model to achieve the desired signal. The work of the writer to Dual-MMSE. Alternatively, as in[42] the performance of the proposed approach using the sampling frequency considered to be 8 kHz, the author will use a realistic 16 kHz sampling rate to perform the analysis.

By using an over-subtraction function (α) and spectral floor (β) in the magnitude field, Naik D C et al [30] suggested an algorithm. In the power spectral domain, as the initial approach suggested by [2] is applied. Similar to the effects of both the applied and proposed approach by subjective and objective performance measures, the researcher has implemented the algorithm suggested by [1] using α and β. The author can conclude that the proposed algorithm gives good results through the results.Nasir Saleem et.al [31] proposed a system consisting of a supervised learning technique to improve a speech-babble-degraded signal. The suggested solution is assembled for speech enhancement with less aggressive Wiener filtering (LW) and deep neural networks (DNNs), known as the DNN-LW. The suggested solution consists of the phases of training and testing. The DNN in the training stage concurrently calculates the magnitude spectrum of noise-free speech and the noise signals from the speech features that are masked by the input noise. To build the improved magnitude spectrum the less aggressive Wiener filter is then placed as an extra layer on top of the deep neural network. Finally, to recover the estimated of a clear speech signal, the phase of a noisy speech signal is used. During the testing stage, to achieve the enhanced speech signal, the trained DNN is provided with the features of noise-masked speech signals. Recently, in speech enhancement applications [40], which are learning techniques with multi-layer representation, deep learning techniques are considered. Because of the non-linear model, each layer converts the representation from one higher level to another. A large training set of 104 noise classes was used to train the network. The various objectives of the binary mask, the perfect ratio mask, the Gamma tone frequency energy range, the STFT spectral magnitude, and the FFT mask were intended to train Wang et al. DNN for speech separation [41].Qizheng Huang et al [32] address a novel approach to speech enhancement using a multi-band excitation (MBE) system based on a deep neural network (DNN). The proposed system generally consists of two levels, namely the level of learning and the stage of development. Two DNNs with different objectives were trained in the training stage. The training goals are the role of the clean speech signal for the harmonic magnitude and band gap. The log-power spectra (LPS) of the noisy speech signal is the input feature of two DNNs. The enhanced speech signal can be obtained through MBE speech synthesis by using the output of DNNs and the estimated online pitch period. Using the proposed approach, the MBE model's parameters can be accurately approximated to synthesize the enhanced speech signal with minimal noise and intelligibility of the high quality speech signal. At the same time, this effectively eliminates the distortion between the harmonics.In the MBE system, the spectrum of speech is split into several sub-bands and a voiced / unvoiced (V / UV) decision is made for each sub-band. Periodic excitation generates the voiced sub-band, while random noise contains the unvoiced sub-band. The MBE model is derived from speech coding, which can produce a speech signal of high quality.



Hanwook Chung et al [33] address how to deal with individual noise characteristics when learning a DNN state for spectral mapping. The author uses noise-dependent adaptation vectors to adjust the weights and biases of the spectral mapping DNN, which are found based on the output of an additional noise classification DNN. During the training stage, the parameters of DNN spectral mapping, DNN noise classification, and vectors for adaptation are calculated jointly. The researcher integrates a classic unsupervised speech enhancement algorithm with the proposed DNN-based approach in the enhancement stage to further improve the improved speech signal quality author validate spectral mapping DNN weights and biases via noise-dependent adaptation vectors. The latter are obtained on the basis of a DNN description of auxiliary noise production. During the training stage, the parameters for spectral mapping DNN, noise detection DNN and the vectors of adaptation are determined together. In addition, the author combines a classic non-supervised speech enhancement algorithm with the DNN-based method to further enhance the quality of speech.Haemin Yang et al [34] used a soft-decision algorithm using the likelihood of speech presence (SPP) to efficiently suppress spectral density (PSD) of noise power This module first measures the possibility of speech in each time frequency bin and then controls the amount of noise update[43],[44]. However, it takes a lot of heuristics to calculate the threshold parameters to find noise PSD with SPP estimation. Therefore, conventional PSD noise-based statistical model estimates require a large number of rule-based experiments to calculate several parameters appropriately, which also does not guarantee high performance. Deep learning architectures are well known to have excellent performance in modeling the non-linear relationship between input and target output data, recently showing significant performance improvements in different research areas. In particular, with an adequate number of noise types and database size, a deep learning-based noise estimator PSD is robust. Deep learning structure, however, is not suitable for real-time communication systems, as it requires a large amount of memory and computational resources to accomplish reasonable performance. This paper focuses on improving the performance of the PSD noise estimation algorithm with a deep SPP estimation system based on training. This paper also reflects on the quality assessment of these two network sizes by setting the network for deep learning technology to two different network sizes. With less complexity, the proposed method offers statistically significant improvement in efficiency.

REFERENCES

[1] S.Boll,"Suppression of Acoustic Noise in Speech Using Spectral Subtraction" IEEE Trans. Acoust, Speech, Signal Process. Vol.27, pp-113-120, Apr 1979.

[2] M.Berouti, R. Schwartz and J.Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise" Proc.IEEE Int Conf Acoust., Speech, Signal Process., pp.208-211,Apr 1979.

[3] Yariv Ephraim and David Malah, "Speech Enhancement Using a- Minimum Mean- Square Error Short-Time Spectral Amplitude Estimator" IEEE Transactions on Acoustics, Speech, and Signal Processing, VOL. ASSP-32, NO. 6, DECEMBER 1984.

[4] Yariv Ephraim,"Statistical Model Based Speech Enhancement Systems" Proceedings of the IEEE, Vol 80, No 10, October 1992.

[5] Olivier Cappe,"Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor" IEEE transactions on speech and audio processing, vol. 2, no. 2, april 1994.

[6] Pascal Scalart and Jozue Vieira filho,"Speech Enhancement Based on a Priori Signal to Noise Estimation", 1996.

[7] Sunil D Kamath and P. Loizou,"A Multi-band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise" in Proceedings Int.Conf. Acoustic, Speech, Signal Processing, Orlando, USA, May 2002, Vol 4, pp. 4160-4164.

[8] Cyril Plapous, Claude Marro, Laurent Mauuary, Pascal Scalart, "A Two- Step Noise Reduction Technique", ICASSP 2004

[9] Y. Hu and P. Loizou,”Evaluation of objective measures for speech enhancement”, in Proc. Interspeech, pp. 14471450, 2006.

[10] Cyril Plapous , Claude Marro, and Pascal Scalart, ”Improved Signal-to- Noise Ratio Estimation for Speech Enhancement” IEEE Transactions on Audio, speech, and Language Processing, VOL. 14, NO. 6, NOVEMBER 2006.

[11] Sundarrajan Rangachari, Philipos C. Loizou,”A noise-estimation algorithm for highly Non-stationary environments”, Science Direct, Speech Communication, 2006, PP-220-231.

[12] Ningping Fan, Justinian Rosca, Radu Balan,”Speech Noise Estimation Using Enhanced Minima Controlled Recursive Averaging”, ICASSP 2007.

[13] Yi Hu and Philipos C. Loizou,”Evaluation of Objective Quality Measures for Speech Enhancement”, IEEE Transactions on Audio, Speech, and Language Processing, VOL. 16, NO. 1, January 2008.

[14] Guangyan Wang, Xia Wang, Xiaoqun Zhao,”Speech Enhancement Based on a Combined Spectral Subtraction with Spectral Estimation in Various Noise Environment”, ICALIP 2008.

[15] Yang Lu, Philipos C. Loizou,”A geometric approach to spectral subtraction”, Science Direct, Speech Communication, and pp: 453-466, 2008.

[16] Jomg Mo Kum, Yun Sik Park and Joon Hyuk Chang,”Speech Enhancement Based on Minima Controlled Recursive Averaging Incorporating Conditional Maximum A Posteriori Criterion”, ICASSP 2009.

[17] Kuldip Paliwal, Kamil Wo jcicki, Belinda Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain”, Science Direct, Speech Communication, pp No: 450-475, 2010.

[18] Hou Xuchu, Zhu Xiaojing,”Speech Enhancement using Harmonic Regeneration”, 2011.

[19] Jesper Rindom Jensen, Jacob Benesty, Mads Grsbll Christensen and Sren Holdt Jensen, “Enhancement of Single-Channel Periodic Signals in the Time-Domain”, IEEE Transactions on Audio, Speech, and Language Processing, VOL. 20, NO. 7, SEPTEMBER 2012.

[20] Navneet Upadhyay and Abhijit Karmakar,”An Improved Multi-Band Spectral Subtraction Algorithm for Enhancing Speech in Various Noise Environments” International Conference on Design and Manufacturing, IConDM 2013.

[21] Martin Krawczyk and Timo Gerkmann,”STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement””, IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 22, NO. 12, DECEMBER 2014.

[22] Josef Kulmer and Pejman Mowlaee,”Phase Estimation in Single Channel Speech Enhancement Using Phase Decomposition”, IEEE Signal Processing Letters, VOL. 22, NO. 5, May 2015.

[23] Yan Xiong, Qiang Chen, Fang Xu and Jun Zhang,”Speech Enhancement Based on Multi-Stream Model”, 6th International Conference on Digital Home, 2016.

[24] Betty Kurian, Shanavaz K T, Nikhil G Kurup, ”PNCC Based Speech Enhancement and Its Performance Evaluation using SNR loss”, 2017 International Conference on Networks & Advances in Computational Technologies (NetACT),Trivandrum 20-22 July .

[25] Qiquan Zhang, Mingjiang Wang,”Speech Enhancement for Nonstationary Noise Environments”, 17th IEEE International Conference on Communication Technology, 2017.

[26] Zhang Wenlu and Peng Hua,”Modified Wiener Filtering Speech Enhancement Algorithm with Phase Spectrum Compensation”, 9th IEEE International Conference on Communication Software and Networks, 2017.

[27] Nahma, L., Yong, P. C., Dam, H. H., & Nordholm, S,”Improved a priori snr estimation in speech enhancement”, 23rd Asia-Pacific Conference on Communications (APCC), 2017.

[28] Siddala Vihari, Dr.A.Sreenivasa Murthy,Priyanka Soni and Naik.D.C, ”Comparison of Speech Enhancement Algorithms” in proc.ICISP ,pp.666-676,Aug 2016.

[29] Julien Basco and Eric Plourde, ”Speech Enhancement Using both Spectral and Spectral Modulation Domains” 2017 IEEE 30th Canadian conference on Electrical and Computer Engineering, CCECE-2017

[30] Naik.D.C, Dr.A.Sreenivasa Murthy and Ramesh Nuttaki,” Modified Magnitude Spectral Subtraction Methods for Speech Enhancement” 2017 international conference on Electrical, Electronics, Communication, Computer and Optimization techniques(ICEECCOT), pp.274-279,2017.

[31] Nasir Saleem, Muhammad Irfan, Xuhui Chen, Muhammad Ali, ”Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise”, IEEE ICIS 2018, June 6-8, 2018, Singapore.

[32] Qizheng Huang, Changchun Bao, Xianyun Wang, Yang Xiang,”DNNBased Speech Enhancement Using MBE Model”, International Workshop on Acoustic Signal Enhancement (IWAENC2018), Sept. 2018, Tokyo, Japan.

[33] Hanwook Chung, Taesup Kim, Eric Plourde and Benoit Champagne, “Noise-Adaptive Deep Neural Network For Single-Channel Speech Enhancement”, 2018 IEEE International Workshop on Machine Learning For Signal Processing, SEPT. 1720, 2018, AALBORG, DENMARK.

[34] Haemin Yang, Soyeon Choe, Keulbit Kim, and Hong-Goo Kang, ”Deep Learning-based Speech Presence Probability Estimation for Noise PSD Estimation in Single-channel Speech Enhancement”, 2018 International Conference on Signals and Systems (ICSigSys), 2018.

[35] L. Cohen and B Berdugo, ”Noise Estimation Based by Minima Controlled Recursive Averaging For Robust Speech Enhancement”, IEEE Signal Processing Letters, vol. 9, no. 1, pp. 12-15, January 2002.

[36] J Benesty and J. Chen, ”Optimal Time-Domain Noise Reduction Filters A Theoretical Study”, Springer Briefs in Electrical and Computer Engineering, 1st Ed. New York: Springer, 2011, no. VII.

[37] M. G. Christensen and A. Jakobsson, ”Optimal filter designs for separating and enhancing periodic signals”, IEEE Trans. Signal Process., vol. 58, no. 12, pp. 59695983, Dec. 2010.

[38] T. Gerkmann, M. Krawczyk, and R. Rehr,”Phase estimation in speech enhancement unimportant,

important, or impossible?" in Proc. IEEE Conv. Elect. Electron. Eng. Israel, Eilat, Israel, Nov. 2012.

[39] I Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands" in Proc IEEE Workshop on Hands Free Speech Communication, HSC01, Kyoto, Japan, Apr. 9-11, 2001, pp. 95-98.

[40] Y. X. Wang and D. L. Wang, "Towards scaling up classification based speech separation," IEEE Trans. on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1381-1390,2013.

[41] Y. Wang, A. Narayanan, and D. Wang, " On training targets for supervised speech separation", IEEE Trans. on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849-1858,2014.

[42] K. Paliwal, K. Wojcicki, and B. Schwerin,"Single-channel speech enhancement using spectral subtraction in the short-time modulation domain", Speech Communication, vol. 52, no. 5, pp. 450475, 2010.

[43] T. Gerkmann and R.C. Hendriks, "Noise power estimation based on the probability of speech presence", in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on. IEEE, 2011, pp. 145148.

[44] A. Hussain, K. Chellappan, and S.M. Zamratol, "Single channel speech enhancement using ideal binary mask technique based on computational auditory scene analysis", Journal of Theoretical and Applied Information Technology, vol. 91, no. 1, pp. 12, 2016.