International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-9506-9

Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing

Thimmaraja G. Yadava[1] · H. S. Jayanna[2]

Received: 16 August 2017 / Accepted: 29 March 2018

Abstract

Speech data collected under uncontrolled environment need to be processed to build a robust automatic speech recognition system. In this paper, a method is proposed to process the degraded speech signal. Initially, the significance of the spectral subtraction with voice activity detection (SS-VAD) and magnitude squared spectrum estimators are studied for different types of noises. In SS-VAD method, the degraded speech data is sampled and windowed into 50% overlapping. The VAD is used to detect the voiced regions of speech signal. The minimum mean square error-short time power spectrum, minimum mean square error-spectrum power based on zero crossing (MMSE-SPZC) and maximum a posteriori estimators are studied individually. These MSS estimators are implemented on the assumption that the magnitude squared spectrum of the degraded speech signal is the sum of the clean (original) speech signal and noise model. The experimental results show that the MMSE-SPZC estimator gives better performance compared to the other two methods. This estimator is combined with SS-VAD method to improve the performance. In this paper, the combined SS-VAD and MMSE-SPZC method, yields better speech quality by reducing noise in degraded speech signal compared to the individual methods.

Keywords Automatic speech recognition (ASR) · Spectral subtraction (SS) voice activity detection (VAD) · Magnitude squared spectrum (MSS) · Speech data

1 Introduction

Speech enhancement mainly depends on the human perceptual factors and signal processing applications. The speech data collected in the real time environment is noisy in nature. Normally speech is corrupted by several degradations such as background noise, vocal noise, factory noise, f16 noise, babble noise and reverberations etc. The noise reduction in degraded speech data is a challenging task (Rabiner and Juang 1993; Loizou 2007). The spectral subtraction (SS) method is commonly used for speech enhancement and is mainly associated with voice activity detection (VAD). To find the active regions of degraded speech signal, VAD is used (Ramirez et al. 2003). The corrupted speech signal is the sum of clean (original) speech signal and additive noise model.

The degraded speech segments are processed frame by frame with a duration of 20 ms. The SS-VAD method was proposed for speech enhancement in Boll (1979), Kamath and Loizou (2002), Jounghoon and Hanseok (2003), Cole et al. (2008) and Goodarzi and Seyedtabaii (2009). The effect of noise can be eliminated in degraded speech signal by subtracting the average magnitude spectrum of noise model from the average magnitude spectrum of degraded speech signal. The process of using several noise elimination techniques for speech enhancement is called speech preprocessing (Loizou 2007). The modified SS algorithm was proposed for speech enhancement in Bing et al. (2009). This algorithm was implemented by using VAD and minima controlled recursive averaging (Cohen and Berdugo 2002). The experimental results are evaluated under ITU-T G.160 standard and compared with existing methods. In Huanhuan et al. (2012), an improved SS

✉ Thimmaraja G. Yadava
thimrajyadav@gmail.com

[1] Department of Electronics and Communication Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

[2] Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

Published online: 10 April 2018

algorithm was proposed for musical noise suppression in degraded speech signal. In that work, the VAD was used for the detection of voiced regions in degraded speech signal. The experimental results show more suppression of musical noise in degraded speech signal.

Various speech signal magnitude squared spectrum (MSS) estimators were proposed for noise reduction in degraded speech signal in Ephraim and Malah (1984, 1985) and Yang and Philipos (2011). The MSS estimators namely, minimum mean square error-short time power spectrum (MMSE-SP), minimum mean square error-spectrum power based on zero crossing (MMSE-SPZC) and Maximum *a posteriori* (MAP) are implemented individually. These MSS estimators significantly performed well under many degraded conditions (Yang and Philipos 2011). Ephraim and Malah have proposed a Minimum Mean Square Error Short Time Spectral Amplitude (MMSE-STSA) estimator for speech enhancement (Ephraim and Malah 1984). This method was compared with most widely used algorithms such as SS and Wiener filtering and it was observed that the proposed MMSE-STSA method gives better performance than the existing methods. An alternatives to the Ephraim and Malah speech enhancement method was proposed under the assumption that the Fourier series expansion of clean (original) speech signal and noise may be modeled as independently with zero mean and Gaussian random variables (Wolfe and Godsill 2001). Martin proposed an algorithm for speech enhancement using MMSE estimators and Supergaussian Priors (Martin 2005). The main significance of this algorithm was to improve the short time spectral coefficients of corrupted speech signal. This method was compared with Wiener filtering and MMSE-STSA methods (Ephraim and Malah 1984). Philipos C. Loizou have proposed an algorithm for noise reduction in corrupted speech signal using Baysian estimators (Philipos 2005). Three different types of Baysian estimators are implemented for speech enhancement.

The literature reveals that the suppression of musical and babble noises efficiently in degraded speech signal is not addressed. Therefore, a method is proposed for the speech enhancement by combining SS-VAD and MMSE-SPZC estimator to suppress the musical, babble and other types of noises in degraded speech signal. The remainder of the paper is organized as follows: The background work and assumptions are given in Sect. 2. The detailed description of SS-VAD is given in Sects. 3 and 4 the derivation of different MSS estimators are given. Section 5 gives the performance measures description. The performance analysis of existing methods is given in Sect. 6. The proposed combined SS-VAD and MMSE-SPZC estimator is given in Sect. 7. The conclusions are given in Sect. 8.

2 Background work and assumptions

Consider a clean (original) speech signal s(n) is corrupted by background noise d(n) which leads to the corrupted speech signal c(n). It can be written as follows:

$$c(t) = s(t) + d(t) \tag{1}$$

The corrupted speech signal is now in time domain which can be converted into frequency domain by sampling at time $t = nT_s$. The resultant corrupted speech signal in frequency domain can be written as follows:

$$c(n) = c(nT_s) \tag{2}$$

where T_s is the sampling duration which can also be written as

$$f_s = \frac{1}{T_s} \tag{3}$$

The short time Fourier transform of c(n) can be written as

$$C(w_k) = S(w_k) + D(w_k) \tag{4}$$

The polar form of the above equation can be shown below.

$$C_k e^{j\theta_c(k)} = S_k e^{j\theta_s(k)} + D_k e^{j\theta_d(k)} \tag{5}$$

where $\{C_k, S_k, D_k\}$ represents the magnitudes and $\{\theta_c(k), \theta_s(k), \theta_d(k)\}$ represents the phase of the noisy speech signal, clean (original) speech signal and noise model respectively. Assuming that the clean (original) speech signal s(n) and noise model d(n) are uncorrelated not moving random processes. The power spectrum of the corrupted speech signal is the sum of the power spectra of clean speech signal and noise model. It can be written as follows:

$$P_c(w) = P_s(w) + P_d(w) \tag{6}$$

Another two assumptions are used in the derivation of MSS estimators. The first assumption is that the power spectrums of clean speech signal, corrupted speech signal and noise model are approximately equal to the magnitude spectrums of clean speech signal, corrupted speech signal and noise model. Therefore, (6) can be written as follows:

$$C_k^2 \approx S_k^2 + D_k^2 \tag{7}$$

The above assumption are usually used in traditional SS algorithms (Boll 1979; Kamath and Loizou 2002; Jounghoon and Hanseok 2003; Cole et al. 2008; Goodarzi and Seyedtabaii 2009; Marc Karam and Hasan 2014). In the remainder of the paper, we will be calling C_k^2, S_k^2 and D_k^2 are as MSSs of corrupted speech signal, clean speech signal and noise model respectively. The second assumption is that the complex part of Discrete Fourier Transform (DFT) coefficients are modeled as free Gaussian random variables.

The probability density functions of S_k^2 and D_k^2 are written as follows:

$$f_{S_k^2} = \frac{1}{\sigma_s^2(k)} e^{-}\frac{S_k^2}{\sigma_s^2(k)} \tag{8}$$

$$f_{D_k^2} = \frac{1}{\sigma_d^2(k)} e^{-}\frac{D_k^2}{\sigma_d^2(k)} \tag{9}$$

where $\sigma_s^2(k)$ and $\sigma_d^2(k)$ can be written as follows.

$$\sigma_s^2(k) \equiv E\{S_k^2\}, \quad \sigma_d^2(k) \equiv E\{D_k^2\} \tag{10}$$

The posterior probability density function of clean (original) speech signal MSS can be computed using natural Bayes theorem as shown below.

$$f_{S_k^2}(S_k^2 \mid C_k^2) = \frac{f_{C_k^2}(C_k^2 \mid S_k^2) f_{S_k^2}(S_k^2)}{f_{C_k^2}(C_k^2)} \tag{11}$$

$$f_{S_k^2}(S_k^2 \mid C_k^2) = \begin{cases} \Psi_k e^{-\frac{S_k^2}{\lambda(k)}} & \text{if } \sigma_s^2(k) \neq \sigma_d^2(k) \\ \frac{1}{C_k^2} & \text{if } \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{12}$$

where $S_K^2 \in \left[0, C_k^2\right]$ and $\lambda(k)$ can be written as follows.

$$\frac{1}{\lambda(k)} \equiv \frac{1}{\sigma_s^2(k)} - \frac{1}{\sigma_d^2(k)} \text{ if } \sigma_s^2(k) \neq \sigma_d^2(k) \tag{13}$$

and

$$\Psi_k \equiv \frac{1}{\lambda(k)\left\{1 - \exp\left[\frac{C_k^2}{\lambda(k)}\right]\right\}} \tag{14}$$

remember if $\sigma_s^2(k) > \sigma_d^2(k)$, then $1/\lambda(k)$ is less than 0 and it is reversible. Hence, Ψ_k in (12) is positive (Yang and Philipos 2011).

3 Spectral subtraction with VAD

SS method is commonly used for noise cancellation in degraded speech signal (Berouti et al. 1979; Etter and Moschytz 1994; Sim et al. 1998; Diethorn 2004; Faller and Chen 2005). The VAD plays an important role in the detection of only voiced area in the speech signal (Ramirez et al. 2003). In this method, we have considered the clean speech signal with 6 s duration and it is corrupted by different noises, includes Additive White Gaussian Noise (AWGN), car, factory and f16 noises. The corrupted speech signal c(n) is converted into segments and each segment consists of 256 samples with the sampling frequency of 8 kHz. The frame overlapping rate of 50% is considered and Hanning window is used in this work. The mathematical representation of Hanning window is as follows:

$$W[n] = 0.5 - 0.5\cos\left(\frac{2\pi n}{N}\right) \quad \text{where } 0 \ \leq n \leq N \tag{15}$$

where N is the number of points and the window length L can be written as $L = N + 1$.

The basic building block diagram of SS-VAD is given in Fig. 1. It consist of several main steps namely, windowing, Fast Fourier Transform (FFT) calculation, noise estimation, half wave rectification, residual noise reduction and calculation of Inverse Fast Fourier Transform (IFFT). The corrupted speech signal c(n) is the combination of clean (original) speech signal s(n) and additive background noise d(n). The corrupted speech signal c(n) is given as an input to the spectral subtracter. The corrupted speech signal is Hanning windowed and the FFT is calculated. The FFT is one of the most important method to analyze the speech spectrum. The active regions of speech signal is identified by VAD (Ramirez et al. 2003), hence the noise is estimated. The linear prediction error (LPE) is mainly associated with Energy E of the signal and zero crossing rate (ZCR). The parameter Y can be written as follows:

$$Y = E(1 - Z)(1 - E) \quad \text{for single frame} \tag{16}$$

$$Y_{\max} = Y \quad \text{for all frames} \tag{17}$$

where Z and L are ZCR and LPE respectively. The fraction term $Y/Y_{\max}$ is used to know whether a signal has voice activity or not. The average magnitude spectrum of VAD output is subtracted with the average magnitude spectrum of noise estimated. Hence this process is called SS with VAD (Boll 1979; Kamath and Loizou 2002; Jounghoon and Hanseok 2003; Cole et al. 2008; Goodarzi and Seyedtabaii 2009; Martin 2001).

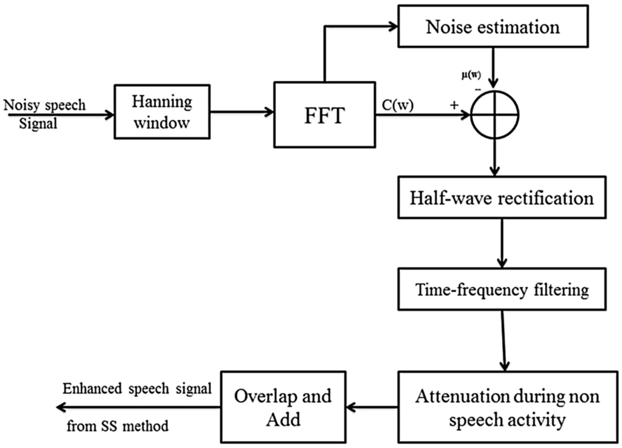

Fig. 1 Block diagram of SS-VAD method

3.1 Role of half wave rectifier (HWR)

The output of SS with VAD can be written as follows.

$$|X_i(w)| = |C_i(w)| - |\mu_i(w)| \tag{18}$$

where $w = 0, 1, 2, ..., L - 1$ and $i = 0, 1, 2, ..., M - 1$. The term L indicates the length of FFT and M indicates the number of frames. The HWR is used in this work to set the spectrums negative values to zero if they have negative values.

3.2 Reduction of residual noise and overlap and add process

Reduction of residual noise in enhanced speech signal is the final step of SS. During the non-speech activities, it is needed to further attenuate the signal. This improves the quality of the enhanced speech signal. Finally, the enhanced speech signal is obtained by calculating its IFFT. The enhanced speech signal can be used in speech processing applications such as, speech recognition, speaker identification, speaker verification, speaker recognition etc.

4 Magnitude squared spectrum estimators

In this work, the following three types of MSS estimators are implemented and their performance is compared.

- Minimum MMSE-SP estimator.
- Minimum MMSE-SPZC.
- MAP estimator.

4.1 Minimum mean square error-short time power spectrum (MMSE-SP) estimator

Wolfe and Godsill (2001) have proposed an algorithm for MMSE-SP estimator. The clean (original) speech signal can be obtained by taking the expected value of clean speech signal and Fourier transform of corrupted speech signal C(w). It can be written as follows:

$$\widehat{S_K^2} = E\{S_k^2|C(w_k)\} \tag{19}$$

$$\widehat{S_K^2} = \int_0^\infty S_k^2 f_{S_k}(S_k|C(w_k)\ dS_k \tag{20}$$

$$\widehat{S_K^2} = \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1+\xi_k}\right)C_k^2 \tag{21}$$

where the terms ξ_k and γ_k represents the a priori and a posteriori SNRs respectively.

$$\xi_k \equiv \frac{\sigma_s^2(k)}{\sigma_d^2(k)}, \quad \gamma_k \equiv \frac{C_k^2}{\sigma_d^2(k)} \tag{22}$$

the implementation steps of this estimator are based on Rician posterior density function $f_{(S_k)}(S_k|Y(w_k))$. It can be represented as follows:

$$f_{S_k}(S_k|C(w_k)) = \frac{S_k}{\sigma_k^2}\exp\left(\frac{S_k^2 + u_k^2}{2\sigma_k^2}\right)I_0\left(\frac{S_k u_k}{\sigma_k^2}\right) \tag{23}$$

where

$$\frac{1}{\lambda'(k)} \equiv \frac{1}{\sigma_s^2(k)} + \frac{1}{\sigma_d^2(k)} \tag{24}$$

$$v_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k \tag{25}$$

$$\sigma_k^2 \equiv \frac{\lambda'(k)}{2} \quad \text{and} \quad u_k^2 \equiv v_k \lambda'(k) \tag{26}$$

where $I_0(\cdot)$ is the 0th order modified Bessel function. The approximate values of Bessel function are calculated in order to derive magnitude spectrums of MAP estimator (Lotter and Vary 2005).

4.2 Minimum mean square error-spectrum power estimator based on zero crossing (MMSE-SPZC)

An another important MSS estimator is MMSE-SPZC. By using (6) and (7), the MMSE-SPZC estimator is derived (Yang and Philipos 2011). The initial MMSE estimator is obtained by calculating the mean of a posteriori density function as shown in (7).

$$\widehat{S_k^2} = E\{S_K^2|C_k^2\} \tag{27}$$

$$\widehat{S_K^2} = \int_0^{C_k^2} S_k^2 f_{S_k^2}(S_k^2|C_k^2)dS_k^2 \tag{28}$$

$$\widehat{S_K^2} = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right)C_k^2, & \text{if } \sigma_s^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2}C_k^2, & \text{if } \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{29}$$

where v_k can be written as

$$v_k \equiv \frac{1-\xi_k}{\xi_k}\gamma_k \tag{30}$$

the estimator gain function can be represented mathematically as follows:

$$G_{MMSE}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right)^{\frac{1}{2}} & \text{if } \sigma_s^2(k) \neq \sigma_d^2(k) \\ \left(\frac{1}{2}\right)^{\frac{1}{2}} & \text{if } \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{31}$$

the gain function of MMSE-SPZC estimator is mainly depends on the parameters ξ_k and γ_k.

4.2.1 Maximum a posteriori (MAP) estimator

The MAP estimator can be represented as follows.

$$\widehat{S_k^2} = \text{argmax} f_{S_K^2}(S_k^2 | C_k^2) \tag{32}$$

maximization with respect to S_K^2.

$$\widehat{S_k^2} = \begin{cases} C_k^2 & \text{if } \frac{1}{\lambda(k)} < 0 \\ 0 & \text{if } \frac{1}{\lambda(k)} > 0 \end{cases} \tag{33}$$

$$\widehat{S_k^2} = \begin{cases} C_k^2 & \text{if } \sigma_s^2(k) \geq \sigma_d^2(k) \\ 0 & \text{if } \sigma_s^2(k) < \sigma_d^2(k) \end{cases} \tag{34}$$

note that the term S_k^2 is bounded in $\left[0, C_k^2\right]$ because of the assumption that the power spectrum is approximating as magnitude spectrum.

The gain function of the MAP estimator can be written as follows:

$$G_{MAP}(k) = \begin{cases} 1 & \text{if } \sigma_s^2(k) \geq \sigma_d^2(k) \\ 0 & \text{if } \sigma_s^2(k) < \sigma_d^2(k) \end{cases} \tag{35}$$

by using (22), the above MAPs gain function can also be represented as:

$$G_{MAP}(\xi_k) = \begin{cases} 1 & \text{if } \xi_k \geq 1 \\ 0 & \text{if } \xi_k < 1 \end{cases} \tag{36}$$

from the above equation we observed that the gain function of MAP estimator is binary in nature. In fact, it is almost same as the binary mask which is widely used in Computational Auditory Scene Analysis (CASA) (2006). The gain function of MAP estimator is based on a priori SNR and the gain function of binary mask is based on instantaneous SNR and this makes a difference between them. The MAP estimator uses the hard thresholding algorithm which can be most widely used in wavelet shrinkage algorithm (Donoho and Johnstone 1995; Jansen 2001; Donoho and Johnstone 1994; Mallat 1999).

5 Performance measures and analysis

The performance of existing methods and proposed method are evaluated from the standard measures. They are Perceptual Evaluation of Speech Quality (PESQ), composite measure and spectrograms described below.

5.1 PESQ

The PESQ measure is an objective measure and it is strongly recommended by ITU-T for quality of speech assessment (Rix et al. 2001; ITU 2000). The term PESQ is calculated as the linear sum of the average distortion value D_{ind} and average asymmetrical distortion value A_{ind}. It can be written as follows (Yi and Philipos 2008):

$$PESQ = b_0 + b_1 D_{ind} + b_2 A_{ind} \tag{37}$$

where $b_0 = 4.5, b_1 = -0.1$ and $b_2 = -0.0309$.

5.2 Composite measures

Composite measures are the objective measures which can be used for the performance evaluation. The ratings and description of different scales are shown in Table 1. The composite measures are derived by multiple linear regression analysis (Hu and Loizou 2006). The multiple linear regression analysis is used to estimate the three important composite measures (Yi and Philipos 2008) they are,

- The composite measure for speech signal distortion (s).
- The composite measure for background noise distortion (b).
- The composite measure for overall speech signal quality (o).

Table 1 The description of the speech signal distortion (s), background noise distortion (b) and overall speech quality (o) scales rating

Ratings	Speech signal scale (s)	Background noise scale (b)	Overall scale (o)
1	Much degraded	Very intrusive and conspicuous	Very poor
2	Fairly degraded and unnatural	Fairly intrusive and conspicuous	Poor
3	Somewhat natural and degraded	Not intrusive and can be noticeable	Somewhat fair
4	Fairly natural with some degradation	Little noticeable	Good
5	Pure natural with no degradation	Can not noticeable	Best and excellent

Table 2 Performance measurement of SS-VAD method in terms of PESQ for TIMIT database

Method	PESQ measure	Musical	Car	Babble	Street
SS-VAD	Input PESQ	1.8569	2.6816	2.3131	1.7497
	Output PESQ	2.1402	3.0823	2.8525	2.2935
	PESQ improvement	**0.2933**	0.4007	0.5394	0.5438

The bold integers indicate the better results compared to other methods

Table 3 Performance measurement of SS-VAD method in terms of PESQ for Kannada database

Method	PESQ measure	Musical	Car	Babble	Street
SS-VAD	Input PESQ	1.8569	2.6816	2.3131	1.7497
	Output PESQ	2.1102	2.9082	2.8625	2.2935
	PESQ improvement	**0.2633**	0.4007	0.5494	0.5438

The bold integers indicate the better results compared to other methods

Table 4 Performance evaluation of SS-VAD method using composite measure for TIMIT database

Method	Composite measure	Musical	Car	Babble	Street
SS-VAD	Speech signal (s)	1.8017	3.6399	3.5213	2.7860
	Background noise (b)	2.6670	2.2760	2.1245	1.9125
	Overall speech quality (0)	**3.1639**	3.7759	3.4245	3.3182

The bold integers indicate the better results compared to other methods

6 Performance analysis of existing methods

The speech was recorded at Texas instruments (TI), transcribed at Massachusetts Institute of Technology (MIT) and verified and prepared for publishing by the National Institute of Standards and Technology (NIST). The TIMIT speech database is used for the conduction of experiments and performance evaluation of existing and proposed methods that are degraded by musical, car, babble and street noises. For local language speech enhancement, Kannada speech database is used and it is also degraded by the same noises respectively. The performances of individual and proposed methods are evaluated as follows:

6.1 Spectral subtraction with VAD results and analysis

The experiments are conducted for TIMIT and Kannada speech databases. The performance measurement of SS-VAD method in terms of PESQ for TIMIT and Kannada databases are as shown in Tables 2 and 3 respectively. From the tables, it was observed that there is a less suppression of noise in degraded speech data which were degraded by musical noise. The SS-VAD is robust in eliminating the noises such as street, babble, car and background noise etc., in corrupted speech data shown in tables. The performance evaluation of SS-VAD method in terms of composite measures is shown in Tables 4 and 5 for TIMIT and Kannada databases respectively. It gives poor speech quality of 3.1639 and 2.9039 for a musical noise compared to other types noises for both databases respectively. Therefore, it is necessary to eliminate the musical noise in degraded speech signal to get good speech quality as like for the speech signals which were degraded by car, babble and street noises.

Table 5 Performance evaluation of SS-VAD method using composite measure for Kannada database

Method	Composite measure	Musical	Car	Babble	Street
SS-VAD	Speech signal (s)	1.9017	3.1199	3.4313	2.2460
	Background noise (b)	2.2370	2.2178	2.1245	1.9355
	Overall speech quality (0)	**2.9039**	3.6659	3.1244	3.2112

The bold integers indicate the better results compared to other methods

6.2 Magnitude squared spectrum estimators results and analysis

In this work, three different types of estimators are implemented. The performance measurement of MMSE-SPZC estimator in terms of PESQ for TIMIT and Kannada speech databases are shown in Tables 6 and 7 respectively. The tables show that there is a much improvement in PESQ for musical, car and street noises compared to babble noise. The poor speech quality is obtained for babble noise after the performance evaluation of the same method using composite measures for both the databases is shown in Tables 8 and 9. Therefore, from the tables it was observed that, the speech sentences were degraded by babble noise should be enhanced efficiently to get good improvement in PESQ as well as good speech quality.

7 Proposed combined SS-VAD and MMSE-SPZC method

The SS-VAD method suppress the various types of noises reasonably such as babble noise, street noise, car noise, vocal noise and background noise etc. The main drawback of SS-VAD is that the suppression of musical noise in degraded speech signal is much less (Loizou 2007; Boll 1979; Cole

Table 6 Performance measurment of MSS estimators in terms of PESQ for TIMIT database

Method	Estimators	PESQ measure	Musical	Car	Babble	Street
MSS estimators	MMSE-SP	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4797	3.3128	2.7043	2.3609
		PESQ improvement	0.6228	0.6312	0.3912	0.6112
	MMSE-SPZC	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4997	3.3337	2.7143	2.3809
		PESQ improvement	**0.6428**	**0.6521**	**0.4012**	**0.6312**
	MAP	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4683	3.2744	2.7129	2.3618
		PESQ improvement	0.6114	0.5928	0.3998	0.6121

The bold integers indicate the better results compared to other methods

Table 7 Performance measurment of MSS estimators in terms of PESQ for Kannada database

Method	Estimators	PESQ measure	Musical	Car	Babble	Street
MSS Estimators	MMSE-SP	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4797	3.3128	2.7043	2.3609
		PESQ improvement	0.6338	0.6113	0.4102	0.6122
	MMSE-SPZC	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4997	3.3337	2.7143	2.3809
		PESQ improvement	**0.6431**	**0.6532**	**0.4101**	**0.6112**
	MAP	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.4683	3.2744	2.7129	2.3618
		PESQ improvement	0.6224	0.5911	0.3998	0.6001

The bold integers indicate the better results compared to other methods

Table 8 Performance evaluation of MSS estimators using composite measure for TIMIT database

Method	Estimators	Composite measure	Musical	Car	Babble	Street
MSS Estimators	MMSE-SP	Speech signal (s)	3.2536	3.8276	5.0913	3.1147
		Background noise (b)	2.3256	2.4908	3.7548	2.0818
		Overall speech quality (o)	3.1002	2.9056	2.0132	2.8925
	MMSE-SPZC	Speech signal (s)	4.5796	3.8336	3.9336	3.1252
		Background noise (b)	3.4031	2.5671	2.1289	2.1211
		Overall speech quality (o)	**4.4565**	**4.2678**	**3.1025**	**4.1815**
	MAP	Speech signal (s)	3.7859	3.6922	3.1563	2.9798
		Background noise (b)	2.8552	2.5627	2.5598	2.1461
		Overall speech quality (o)	3.6478	3.4123	2.8891	3.0814

The bold integers indicate the better results compared to other methods

Table 9 Performance evaluation of MSS estimators using composite measure for Kannada database

Method	Estimators	Composite measure	Musical	Car	Babble	Street
MSS estimators	MMSE-SP	Speech signal (s)	3.2536	3.8276	5.0913	3.1147
		Background noise (b)	2.3256	2.4908	3.7548	2.0818
		Overall speech quality (o)	3.2000	3.1066	2.0132	2.8925
	MMSE-SPZC	Speech signal (s)	4.5796	3.8336	3.9336	3.1252
		Background noise (b)	3.4031	2.5671	2.1289	2.1211
		Overall speech quality (o)	**4.5565**	**4.3679**	**3.2021**	**4.2812**
	MAP	Speech signal (s)	3.7859	3.6922	3.1563	2.9798
		Background noise (b)	2.8552	2.5627	2.5598	2.1461
		Overall speech quality (o)	3.7478	3.5123	2.9099	3.1012

The bold integers indicate the better results compared to other methods

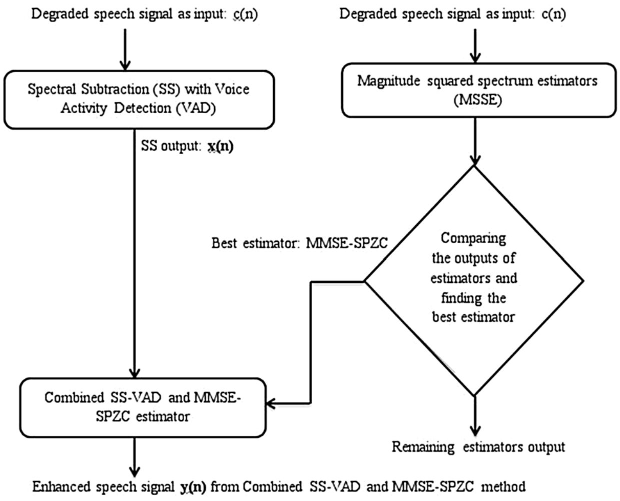
Degraded speech signal as input: c(n)
Degraded speech signal as input: c(n)
Spectral Subtraction (SS) with Voice Activity Detection (VAD)
Magnitude squared spectrum estimators (MSSE)
SS output: x(n)
Best estimator: MMSE-SPZC
Comparing the outputs of estimators and finding the best estimator
Combined SS-VAD and MMSE-SPZC estimator
Remaining estimators output
Enhanced speech signal y(n) from Combined SS-VAD and MMSE-SPZC method

Fig. 2 Flow chart of combined SS-VAD and MMSE-SPZC method

et al. 2008). The MMSE-SPZC is a robust method to suppress the musical noise and given a better results for car, street and white noises compared to babble noise (Yang and Philipos 2011). Therefore, to overcome from the problem of suppression of musical and babble noises, a method is proposed. The proposed method is a combination of the above two methods which suppress the different types of noises including musical and babble noise reasonably under uncontrolled environment. The flowchart of the proposed method is shown in Fig. 2. The output of SS-VAD is little noisier and musical noise is not suppressed as well. Therefore, the output of SS-VAD is passed through MMSE-SPZC estimator.

The MMSE-SPZC estimator reduces the noise in SS-VAD output by considering the low SNR as well as high SNR regions with high intelligibility. The enhanced speech signal from SS-VAD is obtained by subtracting the average magnitude spectrum of noise estimated from the average magnitude spectrum of the speech signal is written in Eq. (18).

The MMSE-SPZC estimator is derived once again for the SS-VAD output. The output x(n) is passed through MMSE-SPZC estimator. Hence the MMSE-SPZC estimator is derived by considering the mean of posteriori density function of SS-VAD output.

$$\widehat{X}_k^2 = E\left\{X_K^2 | Y_k^2\right\} \tag{38}$$

$$\widehat{X}_K^2 = \int_0^{X_k^2} X_k^2 f_{X_k^2}(X_k^2 | Y_k^2) dX_k^2 \tag{39}$$

$$\widehat{X}_K^2 = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right) Y_k^2 & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} Y_k^2 & \text{if } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{40}$$

where X_k and Y_k are the posteriori density functions of SS-VAD output and combined proposed SS-VAD and MMSE-SPZC estimator output respectively. The term v_k is shown in (30).

The gain function of combined SS-VAD and MMSE-SPZC estimator can be written as follows:

Table 10 Performance measurment of combined SS-VAD and MMSE-SPZC estimator in terms of PESQ for TIMIT database

Method	Estimators	PESQ measure	Musical	Car	Babble	Street
Proposed method	SS-VAD and MMSE-SPZC	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.5502	3.3440	3.0912	2.4601
		PESQ improvement	**0.6933**	0.6624	**0.7781**	0.7204

The bold integers indicate the better results compared to other methods

Table 11 Performance measurment of combined SS-VAD and MMSE-SPZC estimator in terms of PESQ for Kannada database

Method	Estimators	PESQ measure	Musical	Car	Babble	Street
Proposed method	SS-VAD and MMSE-SPZC	Input PESQ	1.8569	2.6816	2.3131	1.7497
		Output PESQ	2.5502	3.3440	3.0912	2.4601
		PESQ improvement	**0.7112**	0.6677	**0.7912**	0.7314

The bold integers indicate the better results compared to other methods

Table 12 Performance evaluation of combined SS-VAD and MMSE-SPZC method using composite measure for TIMIT database

Method	Composite measure	Musical	Car	Babble	Street
Proposed method	Speech signal (s)	3.1204	3.6319	3.5689	2.6052
	Background noise (b)	2.6920	2.4780	2.8569	1.9098
	Overall speech quality (0)	**4.4409**	4.3002	**4.2956**	4.3141

The bold integers indicate the better results compared to other methods

Table 13 Performance evaluation of combined SS-VAD and MMSE-SPZC method using composite measure for Kannada database

Method	Composite measure	Musical	Car	Babble	Street
Proposed method	Speech signal (s)	3.1204	3.6319	3.5689	2.6052
	Background noise (b)	2.6920	2.4780	2.8569	1.9098
	Overall speech quality (0)	**4.5111**	4.2911	**4.3123**	4.4112

The bold integers indicate the better results compared to other methods

$$G_{MMSE}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^{v}-1}\right)^{\frac{1}{2}} & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \left(\frac{1}{2}\right)^{\frac{1}{2}} & \text{if } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \quad (41)$$

the gain function is mainly depends on two parameters such as ξ_k and γ_k.

The description of performance measurement of proposed method in terms of PESQ for both the databases is shown in Tables 10 and 11. From the tables, it was observed that there is much suppression in babble and musical noises with PESQ improvements of 0.6933, 0.7781 and 0.7112, 0.7314 for TIMIT and Kannada databases by proposed method compared to individual methods. The speech quality is much improved after the performance evaluation of proposed method using composite measures for both the databases is shown in Tables 12 and 13. From the experimental results and analysis it can be inferred that the combined SS-VAD and MMSE-SPZC method reduces the noise in degraded speech data significantly compared to the individual methods. The enhanced speech data obtained from the proposed method is better audible and high quality than the individual methods. Therefore, the proposed method can be used for Kannada speech database enhancement. The majority of collected speech data is degraded by babble, musical and street noises since it is collected under uncontrolled environment. The spoken query system to access the real time agricultural commodity prices and weather information in Kannada language/dialects is developed for noisy speech data (Thimmaraja and Jayanna 2017). The proposed speech enhancement algorithm could be used for the speech enhancement of collected degraded speech data which was used in Thimmaraja and Jayanna (2017). Therefore, the speech recognition accuracy could be improved in spoken query system by enhancing the noisy speech data and building automatic speech recognition (ASR) models for enhanced speech data.

8 Conclusions

In this work, SS-VAD method is implemented and analyzed its performance. The different MSS estimators are also studied and found that the MMSE-SPZC estimator performs better compared to other estimators. To get better performance for the speech data degraded mainly by musical and babble noises, SS-VAD and MMSE-SPZC combination is proposed. The conducted experimental results show that the proposed method gives better results with high intelligibility and speech quality for the speech data degraded by different types of noises compared to the individual methods. The future challenging work is to apply the proposed method to enhance the speech data collected under uncontrolled environment. After speech enhancement, further work is to build an ASR models for enhanced speech data for the development of spoken query system to access the agricultural commodity prices and weather information in Kannada language/dialects.

References

Beh, J., & Ko, H. (2003). A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech. In *IEEE international conference on multimedia and expo, New York* (Vol. 3, pp. I-648–I-651).

Berouti, M., Schwartz M., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proceedings of IEEE International conference on acoustics, speech and signal processing*, Washington DC (pp. 208–211).

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics Speech and Signal Processing*, *27*, 113–120.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, *120*(6), 4007–4018.

Cohen, I., & Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, *9*(1), 12–15.

Cole, C., Karam, M., & Aglan, H. (2008). Spectral subtraction of noise in speech processing applications. In *40th Southeastern symposium system theory* (pp. 50–53), SSST-2008, 16–18 March.

Computational Auditory Scene Analysis (CASA). (2006). In D. Wang & G. Brown (Eds.), *Principles, algorithms, and applications*. Piscataway, NJ: Wiley/IEEE Press.

Diethorn, E. J. (2004). Subband noise reduction methods for speech enhancement. In Y. Huang & J. Benesty (Eds.), *Audio signal processing for next-generation multimedia communication systems* (pp. 91–115). Boston: Springer.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*(3), 425–455.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

Ephraim, Y., & Malah, D. (1984). 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.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics Speech and Signal processing*, *33*(2), 443–445.

Etter, W., & Moschytz, G. S. (1994). Noise reduction by noise-adaptive spectral magnitude expansion. *Journal of the Audio Engineering Society*, *42*, 341–349.

Evans, N. W. D., Mason, J. S., Liu, W. M., & Fauve, B. (2005). On the fundamental limitations of spectral subtraction: An assessment by automatic speech recognition. In *Signal processing conference, 2005 13th European, Antalya* (pp. 1-4).

Faller, C., & Chen, J. (2005). Suppressing acoustic echo in a spectral envelope space. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1048–1062.

Gauvain, J. L., & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–299.

Goodarzi, H. M., & Seyedtabaii, S. (2009). Speech enhancement using spectral subtraction based on a modified noise minimum statistics estimation. In *Fifth joint international conference* (pp. 1339–1343), August 25–27, 2009.

Hu, Y., & Loizou, P. (2006). Subjective comparison of speech enhancement algorithms. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, Toulouse (Vol. 1, pp. 153–156).

Hu, Y., & Loizou, P. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*, 588–601.

Hu, Y. & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(1), 229–238.

ITU. (2000). *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone net- works and speech codecs*. ITU, ITU-T Rec.

Jansen, M. (2001). *Noise reduction by wavelet thresholding*. Series lecture notes in statistics (Vol. 161). Berlin: Springer.

Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, Orlando

Karam, M., Khazaal, H. F., Aglan, H., & Cole, C. (2014). Noise removal in speech processing using spectral subtraction. *Journal of Signal and Information Processing*, *5*(2), 45989.

Kim, G., & Loizou, P. C. (2010). Improving speech intelligibility in noise using environment-optimized algorithms. *IEEE Transactions on Audio Speech and Language Processing*, *18*(8), 2080–2090.

Kim, G., Lu, Y., Hu, Y., & Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, *126*(3), 1486–1494.

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

Liu, H., Yu, X., Wan, W., & Swaminathan, R. (2012). An improved spectral subtraction method. In *International conference on audio, language and image processing (ICALIP)*, Shanghai (pp. 790–793).

Loizou, P. C. (2005). Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 857–869.

Loizou, P. (2007). *Speech enhancement: Theory and practice* (1st ed.). Boca Raton, FL: CRC Taylor & Francis.

Lotter, T., & Vary, P. (2005). Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, *5*(1), 1110–1126.

Lu, Y., & Loizou, P. C. (2011). Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. *IEEE Transactions on Audio Speech and Language Processing*, *19*(5), 1123–1137.

Mallat, S. (1999). *A wavelet tour of signal processing*. San Diego, CA: Academic.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, *9*(5), 504–512.

Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 845–856.

McAulay, R., & Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics Speech and Signal Processing*, *28*(2), 137–145.

Quackenbush, S., Barnwell, T., & Clements, M. (1988). *Objective measures of speech quality*. Englewood Cliffs, NJ: Prentice-Hall.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Upper Saddle River, NJ: Prentice- Hall Inc.

Ramirez, J., Gorriz, J. M., Segura, J. C., et al. (2003). *Voice activity detection. Fundamentals and speech recognition system robustness*. Rijeka: InTech.

Rix, A., Beerends, J., Hollier, M., & Hekstra, A. (2001). Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, Istanbul (Vol. 2, pp. 749–752).

Sim, B. L., Tong, Y. C., Chang, J. S., & Tan, C. T. (1998). A parametric formulation of the generalized spectral subtraction method. *IEEE Transactions on Speech and Audio Processing*, *6*(4), 328–337.

Thimmaraja, Y. G., Jai Prakash, T. S., & Jayanna, H. S. (2015). Noise elimination in degraded Kannada speech signal for speech recognition. In *IEEE proceedings of international conference on trends in automation, communication and computing technologies (ITACT-2015)*, Bangalore (pp. 183–186), December 21–22, 2015.

Wolfe, P. J., & Godsill, S. J. (2001). Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In *Proceedings of the 11th IEEE signal processing workshop on statistics and signal processing*, Singapore (pp. 496–499).

Xia, B., Liang, Y., & Bao, C. (2009). A modified spectral subtraction method for speech enhancement based on masking property of human auditory system. In *International conference on wireless communications signal processing, WCSP*, Nanjing (pp. 1–5).

Yadava, T. G., & Jayanna, H. S. (2017). A spoken query system for the agricultural commodity prices and weather information access in Kannada language. *International Journal of Speech Technology, Springer*, *20*(3), 635–644.