



# Enhancements in Continuous Kannada ASR System by Background Noise Elimination

G. Thimmaraja Yadava<sup>1</sup> · B. G. Nagaraja<sup>2</sup> · H. S. Jayanna<sup>3</sup>

Received: 22 October 2020 / Revised: 15 January 2022 / Accepted: 18 January 2022 /  
Published online: 16 February 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

In this work, we demonstrate the current advancements assimilated in the earlier developed continuous Kannada automatic speech recognition (ASR) spoken query system (SQS) under uncontrolled environment. The SQS comprises interactive voice response system and ASR models which are developed using Kaldi. A variety of background noises were added to the continuous Kannada speech data while training the ASR system, as it was gathered under a corrupted environment. In the earlier SQS, the background and other types of noises have reduced the accuracy of speech recognition. This can be overcome by developing a robust noise reduction algorithm for degraded speech enhancement. In the enhanced SQS, a background noise reduction module is introduced before the speech feature extraction step. The proposed noise cancellation algorithm is represented by the degraded spectrum of speech in a complex plane which is an amalgamation of clean speech spectrum and noise model vectors. The conducted investigational results reveal that the proposed noise suppression algorithm outperforms the traditional spectral subtraction algorithms and magnitude squared spectrum (MSS) estimators. The outputs of the proposed approach show that there is no audibility of musical noise and other types of noises in enhanced NOIZEUS speech corpora

---

B. G. Nagaraja and H. S. Jayanna have contributed equally.

---

✉ G. Thimmaraja Yadava  
thimrajyadav@gmail.com

B. G. Nagaraja  
nagarajbg@gmail.com

H. S. Jayanna  
jayannahs@gmail.com

<sup>1</sup> E&CE, Nitte Meenakshi Institute of Technology, Yelahanka, Bengaluru, Karnataka 560064, India

<sup>2</sup> E&CE, K.L.E. Institute of Technology, Opposite to Airport, Gokul Road, Hubballi, Karnataka 580027, India

<sup>3</sup> IS&E, Siddaganga Institute of Technology, B. H. Road, Tumkur, Karnataka 572103, India

and continuous Kannada speech data. Therefore, the noise suppression algorithm is applied to the degraded continuous Kannada speech data for its enhancement. Using noise suppression algorithm and time delay neural network ASR modelling technique in SQS, there is an improvement of 1.87% in terms of word error rate in comparison with the earlier developed deep neural network - hidden Markov model (DNN-HMM)-based SQS. The online testing of enhanced continuous Kannada SQS is done by the 500 speakers/users of the Karnataka state under a corrupted environment. The source code of algorithms and ASR models used in this work is made publicly available <https://sites.google.com/view/thimmarajavadav/downloads>.

**Keywords** Automatic speech recognition (ASR) · Interactive voice response system (IVRS) · Spoken query system (SQS) · Continuous Kannada speech data

## 1 Introduction

The speech data gathered under corrupted conditions need to be preprocessed to achieve better advancements in any automatic speech recognition (ASR) systems or spoken query systems (SQS) [25]. As presented in [17], the continuous Kannada speech data were used for system training and decoding. The trained data with Kaldi constituted a high level of background noises that minimized the accuracy of speech recognition. To solve this problem, in this work, we have introduced a noise suppression algorithm at the front step of SQS. The proposed noise elimination algorithm is a spatial procedure (SP) to spectral subtraction (SS) algorithms that have significantly reduced various types of noises in the collected continuous Kannada degraded speech data and has improved the performance of current SQS compared to earlier SQS. The SS technique is one of the most important types of techniques used traditionally for speech enhancement [3, 16, 29]. The SS algorithm involves the approximation of enhanced/processed speech signal spectrum computed by subtracting the approximation of noise spectrum from the spectrum of corrupted speech data. The noise spectrum is computed based on the speech absence, and it is updated for every frequency bin. The process of SS is to be done aptly because too much subtraction leads to loss of speech components and less subtraction may reflect much presence of noise in enhanced speech data [21, 22, 33].

The work in [7] described a noise suppression algorithm using spectral subtraction. The technique utilizes a noise and speech-dependent gain for every frequency section. Further, to reduce the variance of the gain function, spectrum-dependent adaptive averaging was presented. Experimental results showed 10 dB background noise reduction for different SNR conditions (−6 dB to 16 dB). Also in comparison with the SS method, the proposed method achieved improved speech quality and reduced noise artefacts. Overestimation of a spectrum of noise, flooring of the spectrum, dividing the spectrum into few bands of frequency and subjecting to nonlinear methods for each band were presented in [2, 13, 19]. The work in [28] gives some psycho-acoustic rules to adjust the extreme subtraction specifications to render the residue inaudible. Though the SS algorithm reduces noise significantly, it has a major drawback that the subtracted output might have contained some negative values or negative merits due to

errors in approximating the spectrum of noise. The simpler approach to this problem is to set all negative information values to zero and obtain the non-negative spectrum.

Sometimes the musical noise will be having more impact than other types of noises on the listeners. The assumption made in the derivation of the SS algorithm is that the cross-values of the phase differences amongst the noise and clean speech signals are zero. This assumption is considered strong because the clean speech data are not correlated with the introduced noise. The end equations obtained from the SS algorithm are not perfect, but they are approximations. Some of the efforts were made in [6, 14, 35] on how to compensate the cross-terms in the SS algorithm. The performance of compressive sensing (CS)-based method for speech enhancement was studied in [15]. Further, the performances of greedy algorithms, viz., orthogonal matching pursuit, matching pursuit, compressive sampling matching pursuit, stage-wise orthogonal matching pursuit and generalized orthogonal matching pursuit, were compared for speech enhancement. The results were analysed using composite objective measures, and simulation time showed the CS-based technique using generalized orthogonal matching pursuit algorithm achieved better performance than the other recovery algorithms.

The advancements in the Kannada ASR system by background noise elimination and acoustic modelling techniques were described in [32]. The authors have proposed an algorithm, which is a combination of SS–voice activity detection (VAD) and minimum mean square error (MMSE)-spectrum power estimator based on zero-crossing. The experimental results showed that there were significant advancements in terms of speech recognition accuracy for enhanced speech data compared to degraded speech data. Agricultural commodity prices and weather information access based on end-to-end speech recognition system for Kannada dialect were described in [30]. The developed ASR system used noisy Kannada speech data for system training and decoding using Kaldi. The implemented system was designed using an interactive voice response system (IVRS) call flow. The experimental results showed that the word error rate (WER) of online and offline exactly matched with each other. A robust speaker recognition method based on the amalgamation of time-delay neural network (TDNN) and long short-term memory with recurrent project layer (LSTMP) model was described in [18]. The experiments were investigated on four speech corpora. The results showed that the combination of TDNN and LSTMP outperforms the baseline system (*i*-vector).

The work in [27] developed an online Assamese spoken query system for accessing the price of agricultural commodities. Training data, as well as query speech data, includes lots of background noise as the speech data are collected using real farmers. To mitigate the ill effects of the background noise, a zero frequency filtering-based foreground speech separation front-end noise removal scheme was introduced into the ASR system. It was observed an absolute reduction of 6.24% WER is achieved in comparison with the previously reported spoken query system performance [26]. In this work, based on the spatial principles of speech signals, we build a new procedure for the SS algorithm called spatial procedure (SP). The proposed algorithm is based on spatial mathematical procedural steps; henceforth, we will refer to it as the SP technique to the SS algorithm. The proposed algorithm overcomes two drawbacks of the SS algorithm [1]. They are musical noise suppression and false assumptions on

cross-terms are being zero. The proposed SP algorithm is represented by the degraded spectrum of speech in the complex plane is the amalgamation of clean speech data and noise model vectors. The remainder of the paper is organized as follows: Section 2 gives the implementation of the proposed noise elimination technique in detail. The development of a robust enhanced continuous Kannada ASR system is presented in Sect. 3. Section 4 gives the conclusions.

## 2 Background Noise Suppression by SP Technique

### 2.1 Background and Error Analysis for SS

Consider the noisy speech data,  $i(n)$  is the summation of clean or original speech data,  $j(n)$  and noise model,  $l(n)$ . The Fourier transform (FT) of  $i(n)$  can be shown as:

$$I(w_k) = J(w_k) + L(w_k) \quad (1)$$

where  $w_k = \frac{2\pi k}{N}$ ;  $k = 0, 1, \dots, N - 1$  and  $N$  is considered as length of frame. To get the short time power spectra, we have to take the product of  $I(w_k)$  and its conjugate. Therefore, Eq. 1 becomes

$$\begin{aligned} |I(w_k)|^2 &= |J(w_k)|^2 + |L(w_k)|^2 + J(w_k) \cdot L^*(w_k) + L(w_k) \cdot J^*(w_k) \\ &= |J(w_k)|^2 + |L(w_k)|^2 + 2|J(w_k)||L(w_k)|\cos(\theta_J(k) - \theta_L(k)) \end{aligned} \quad (2)$$

The notations  $\{|L(w_k)|^2 + J(w_k) \cdot L^*(w_k) + L(w_k)\}$  and  $L(w_k) \cdot J^*(w_k)$  could not be computed directly and those can be approximated as  $E\{|L(w_k)|^2\}$ ,  $E\{J(w_k) \cdot L^*(w_k)\}$  and  $E\{L(w_k) \cdot J^*(w_k)\}$  where the term  $E\{\cdot\}$  is called as expectation operator. Normally, the operator  $E\{|L(w_k)|^2\}$  is calculated during speech absence and it is described as  $|\widehat{L}(w_k)|^2$ . As per the fundamental assumption, the clean speech signal and noise models are nowhere correlated. Therefore,  $E\{J(w_k) \cdot L^*(w_k)\}$  and  $E\{L(w_k) \cdot J^*(w_k)\}$  are completely equal to zero. Therefore, from the above assumption, the clean speech spectrum estimate can be written as  $|\widehat{J}(w_k)|^2$  and it can be represented as follows:

$$|\widehat{J}(w_k)|^2 = |I(w_k)|^2 - |\widehat{L}(w_k)|^2 \quad (3)$$

Equation 3 shows the traditional power SS technique. The values in  $|\widehat{J}(w_k)|^2$  might be negative. Therefore, the negative values could be made to zero using half wave rectification. The enhanced or processed speech signal can be obtained by taking inverse FT of  $|\widehat{J}(w_k)|$  using phase or angle of the corrupted speech data. Equation 3 can also be represented as:

$$|\widehat{J}(w_k)|^2 = H^2(w_k) \cdot |I(w_k)|^2 \quad (4)$$

and

$$H(w_k) = \sqrt{1 - \frac{|\widehat{L}(w_k)|^2}{|I(w_k)|^2}} = \sqrt{\frac{\gamma(k) - 1}{\gamma(k)}} \quad (5)$$

is the function of gain and  $\gamma(k) \triangleq |I(w_k)|^2 / |\widehat{L}(w_k)|^2$ . The range of  $H(w_k)$  is always  $0 \leq H(w_k) \leq 1$ . There is no compulsory rule for the cross-terms to be zero; sometimes, they will be having large values related to  $|I(w_k)|^2$ . To evaluate the error established from Eq. 3 when the cross-terms are left out, Eq. 2 can be rewritten as:

$$\begin{aligned} |I(w_k)|^2 &= |J(w_k)|^2 + |L(w_k)|^2 + \Delta I(w_k) \\ &= |\widehat{I}(w_k)|^2 + \Delta I(w_k) \end{aligned} \quad (6)$$

where the term,  $|\widehat{I}(w_k)|^2 = |J(w_k)|^2 + |L(w_k)|^2$  and  $\Delta I(w_k)$  defines cross-terms. By ignoring these cross-terms from the above equation, we can write the expression for relative-error as follows:

$$\epsilon(k) \triangleq \frac{|\Delta I(w_k)|}{|I(w_k)|^2} \quad (7)$$

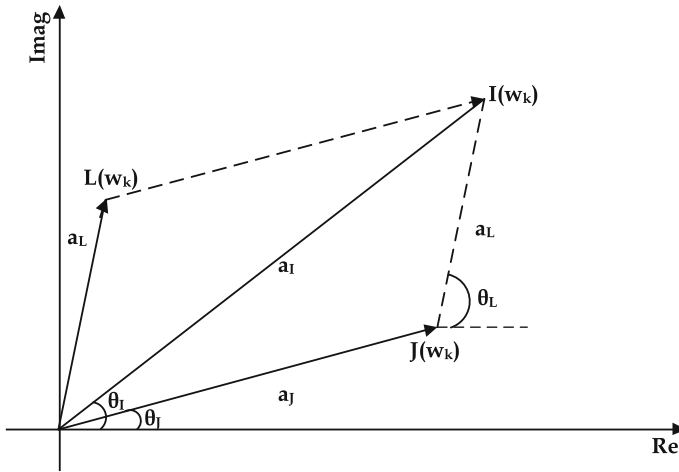
The term  $\epsilon(k)$  is normalized when we considered the spectrum power of degraded speech. The assumption made in Eq. 7 that the spectrum of noise is known factor. The normalized error of cross-terms can be shown in terms of SNR as follows:

$$\epsilon(k) = \left| \frac{2\sqrt{\xi(k)} \cos(\theta_J(k) - \theta_L(k))}{1 + \xi(k) + 2\sqrt{\xi(k)} \cos(\theta_J(k) - \theta_L(k))} \right| \quad (8)$$

where the term  $\xi(k)$  can be written as:  $\xi(k) \triangleq |J(w_k)|^2 / |L(w_k)|^2$  represents exact SNR in frequency bin  $k$ . If  $\xi(k) = 0$ , only when  $\cos(\theta_J(k) - \theta_L(k)) = 0$ , which is equivalent with Eq. 2. From Eq. 8, we could draw some conclusions that, if  $\xi(k) \rightarrow \infty$  or  $\xi(k) \rightarrow 0$ , then  $\epsilon(k) \rightarrow 0$ . Therefore, as the SNR  $\rightarrow \pm\infty$ , we can easily make an assumption that the cross-terms are ignorable. If the SNR values fall in between the range, then the cross-terms could not be neglected. In summary, the error analysis done in this section reveals that the assumption in Eq. 3, the values of cross-terms are zero, is not applicable for spectral SNRs values which are nearly equal to 0 dB. In fact, Eq. 3 can be used in greater errors estimation. Therefore, the error of cross-term  $\epsilon(k)$  is greatest for level of SNR nearly equal to 0 dB and it is consistent with the predictions in [13]. In the next section, we describe a new approach that makes no such assumptions on cross-terms being zero in Eq. 2.

## 2.2 A Spatial Procedure to SS

We propose an algorithm called SP to SS technique which suppresses significant amount of different types of background noises and musical noise in degraded speech data. Using Eq. 1, the degraded speech spectrum  $I(w_k)$  at the frequency  $w_k$  is the



**Fig. 1** The visual representation of degraded speech, noise and clean speech spectrum in complex plane amalgamation of complex valued spectrum of  $J(w_k)$  and  $L(w_k)$ . The complex representation of degraded speech spectrum, noise model spectrum and clean speech signal spectrum in complex plane is shown in Fig. 1. The gain function of SS algorithm is obtained after the assumption that the cross-terms are equal to zero or the angle or phase difference between  $(\theta_J(k) - \theta_L(k))$  is equal to  $\pm\pi/2$  from Eq. 5. In this paper, we derive a formula of generic gain function for SS algorithm which does not depend on any cross-terms and the derived formula will squash the presumptions about the phase difference values between the clean speech signal and noise model. Equation 1 can be rewritten in polar form as follows:

$$a_I e^{j\theta_I} = a_J e^{j\theta_J} + a_L e^{j\theta_L} \tag{9}$$

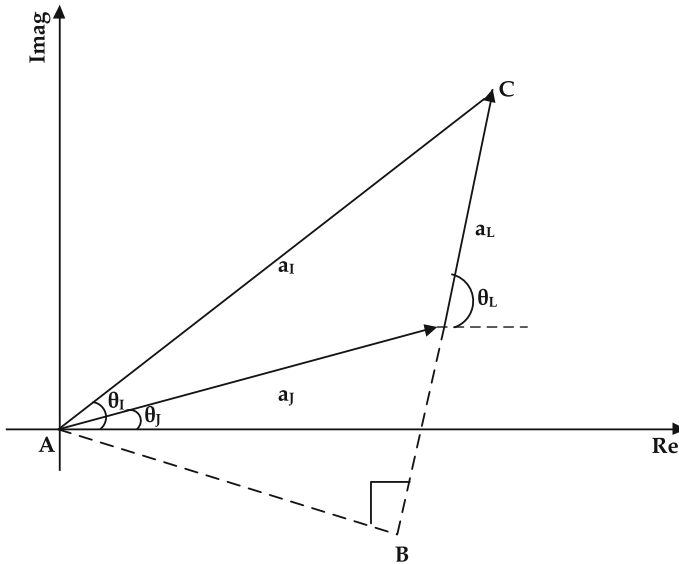
The terms,  $a_I$ ,  $a_J$  and  $a_L$  are the magnitudes and phases are  $\theta_I$ ,  $\theta_J$  and  $\theta_L$  for degraded speech, clean or original speech data and noise model spectrum, respectively.

Figure 2 shows the triangle representation of the spatial relationship amidst the angles of degraded speech signal, noise model and clean or original speech spectrum. From Fig. 2, consider the right angle triangle ABC,  $\overline{AB} \perp \overline{BC}$

$$\begin{aligned} \overline{AB} &= a_I \sin(\theta_L - \theta_I) = a_J \sin(\theta_L - \theta_J) \\ &\Rightarrow a_I^2 \sin^2(\theta_L - \theta_I) = a_J^2 \sin^2(\theta_L - \theta_J) \\ &\Rightarrow a_I^2 [1 - \cos^2(\theta_L - \theta_I)] = a_J^2 [1 - \cos^2(\theta_L - \theta_J)] \\ &\Rightarrow a_I^2 (1 - C_{IL}^2) = a_J^2 (1 - C_{JL}^2) \end{aligned} \tag{10}$$

where  $C_{IL} \triangleq \cos(\theta_I - \theta_L)$  and  $C_{JL} \triangleq \cos(\theta_J - \theta_L)$ . From Eq. 10, we can obtain a new gain function called  $H_{SP}$ . It can be written as follows:

$$H_{SP} = \frac{a_J}{a_I} = \sqrt{\frac{1 - C_{IL}^2}{1 - C_{JL}^2}} \tag{11}$$



**Fig. 2** The spatial relationship amidst the angles of corrupted speech, clean speech data and noise spectrum

The proposed gain function  $H_{SP}$  is always +ve and real because the values of  $C_{IL}$  and  $C_{JL}$  are bounded by 1. The traditional power SS algorithm’s gain function in Eq. 5 is also +ve but smaller than 1, but the proposed gain function can be greater than 1, if the magnitude of  $C_{IL}$  less than  $C_{JL}$ . The suppression from the proposed gain function minimizes to the suppression function of power SS algorithm (from Eq. 5), if  $C_{JL} = 0$ , that is if the speech data vector and noise vector are perpendicular to each other. Cardinaly, the speech signal and the noise signal are perpendicular to each other and are having zero mean, which shows that they are also not correlated with each other [19]. To prove that, the above suppression gain function minimizes to that of Eq. 5 when  $C_{JL}$  is equal to zero. From Fig. 1, we can observe that when the clean speech signal and noise model are perpendicular to each other, then

$$C_{IL} = \frac{a_L}{a_I} \tag{12}$$

Equating the above relation in Eq. 11, we get Eq. 5. Therefore, we can conclude that the suppression mathematical model given in Eq. 11 is true and fact suppression procedure for SS techniques if none of the assumptions were made on the relationship amongst the clean speech signal and noise model. The suppression rule given in Eq. 5 is an approximation which works better over short intervals of time (20–30 ms) by presuming that the original or clean speech and noise model vectors are perpendicular to each other. The product of noisy speech spectrum with suppression function in Eq. 5 could not give the clean speech magnitude spectra, though we have access to the correct spectrum of noise model magnitude. In contrast, the product of noisy speech spectrum with the proposed gain function will exactly give the clean or original speech signal magnitude spectrum shown in Eq. 11. The mentioned suppression gain function

majorly depends on the approximation of the phase or angle differences amongst degraded (or clean) and noise signal models. However, it is a difficult activity and no other techniques currently exist to calculate the values of these phases or angles exactly. One possible solution is to implement and exploit the relationships amidst the angles of noisy or corrupted speech data and noise signals using trigonometric functions. Therefore, we can easily give the solution explicitly for  $C_{IL}$  and  $C_{JL}$  which gives (proof from [28], Appendix B)

$$C_{IL} = \frac{a_I^2 + a_L^2 - a_J^2}{2a_I a_L} \quad (13)$$

$$C_{JL} = \frac{a_I^2 - a_J^2 - a_L^2}{2a_J a_L} \quad (14)$$

The major drawback of the above-mentioned equations is that they completely depend on clean or original speech signal amplitude which we have not considered or which we don't have. Therefore, we need to implement mathematical equations for  $C_{JL}$  and  $C_{IL}$  by dividing both denominator and numerator of Eqs. 13 and 14 by the factor  $a_L^2$ . We get

$$C_{IL} = \frac{\gamma + 1 - \xi}{2\sqrt{\gamma}} \quad (15)$$

$$C_{JL} = \frac{\gamma - 1 - \xi}{2\sqrt{\xi}} \quad (16)$$

where the term  $\xi$  and  $\gamma$  are defined as

$$\xi \triangleq \frac{a_J^2}{a_L^2} \quad (17)$$

$$\gamma \triangleq \frac{a_I^2}{a_L^2} \quad (18)$$

The symbols  $\xi$  and  $\gamma$  are the instantaneous varieties of *a priori* and *posteriori* SNRs, respectively, which were used in traditional minimum mean square error (MMSE) techniques [14, 35]. By equating Eqs. 15 and 16 in Eq. 11, we get the updated proposed function of gain:

$$H_{SP}(\xi, \gamma) = \sqrt{\frac{1 - \frac{(\gamma+1-\xi)^2}{4\gamma}}{1 - \frac{(\gamma-1-\xi)^2}{4\xi}}} \quad (19)$$

The updated gain function in Eq. 19 is approximately equal to the gain functions of MMSE algorithms, and it mainly depends on  $\xi$  and  $\gamma$ . The performance of the gain function in Eq. 19 provides equal suppression with respect to MMSE algorithms under



$(\gamma - 1) < 5$  dB. For values of  $(\gamma - 1) > 5$  dB, the gain function suppression of the SP algorithm becomes more reliable than the MMSE estimators gain functions [28].

In summary, two comparative differences are there with the proposed SP technique and traditional MMSE algorithm. Firstly, the proposed SP algorithm is not random, is deterministic and not implemented using any statistic models. In addition, none of the assumptions were made on the statistical distributions of noisy or clean speech signal and noise FT coefficients, as worked in the case of MMSE techniques. Secondly, the parameters  $\gamma$  and  $\xi$  are instantaneous SNR values. In the next section, we implement the equations for the parameters,  $\gamma$  and  $\xi$ . We evaluate and contrast the reliability of the proposed SP technique using both the instantaneous and long-term frames mean measurements of  $\gamma$  and  $\xi$ .

### 2.3 Implementation

The computation of gain function,  $H_{SP}(\xi, \gamma)$  mainly depends on the estimation of  $\gamma$  and  $\xi$ . As per the definition in Eqs. 17 and 18, the terms,  $\gamma$  and  $\xi$  are the instantaneous SNR values, not long-term and average statistical values [5, 20, 21, 34]. In [5], the expectation operators were used for computing the values of  $\gamma$  and  $\xi$  yielded less performance. Therefore, the methods used in [5] cannot be used to approximate the parameters  $\gamma$  and  $\xi$ . Hence, we propose a new method to estimate the parameters by considering the present and past spectral characteristics information. The expected value of  $\xi$  is calculated by considering the magnitude spectrum of past frames as shown below.

$$\widehat{\xi}_I(\lambda, k) = \widehat{a}_J^2(\lambda - 1, k) / \widehat{a}_L^2(\lambda - 1, k) \quad (20)$$

The term  $\widehat{\xi}_I(\lambda, k)$  is an estimation of  $\xi$  at the frame  $\lambda$  and frequency bin  $k$ , and  $I$  denotes the instantaneous SNR measurement. The above estimate uses only the past spectral characteristic information values. Therefore by amalgamating the two approximations of  $\xi$  computed using present and past spectral characteristic information values, we obtain

$$\widehat{\xi}(\lambda, k) = \alpha \cdot \frac{\widehat{a}_J(\lambda - 1, k)^2}{\widehat{a}_L(\lambda - 1, k)^2} + (1 - \alpha) \cdot (\sqrt{\widehat{\gamma}(\lambda, k)} - 1)^2 \quad (21)$$

The smoothing parameter is denoted by  $\alpha$  in the above equation, and  $\widehat{a}_L(\lambda, k)$  is noise magnitude spectrum estimate. Equation 21 is the weighted mean of present and past SNR values, and  $\alpha$  manages the weight assigned on present and past spectral characteristics information. Equation 21 gives the weighted mean of present and past SNR values, and this approach is approximately equal to decision directed approach used in [5]. If the value of smoothing parameter is equal to 1, then Eq. 21 becomes the estimate of instantaneous values  $\xi$  given in Eq. 20. Similarly, the value of  $\widehat{\gamma}_I(\lambda, k)$  can be shown as follows:

$$\widehat{\gamma}_I(\lambda, k) = \left( \frac{a_I(\lambda, k)}{\widehat{a}_L(\lambda, k)} \right)^2 \quad (22)$$

The term  $\widehat{a}_L(\lambda, k)$  is noise spectrum estimate obtained using estimation of noise technique. The smoothing factor related to  $\widehat{\gamma}_I(\lambda, k)$  is shown in Eq. 23:

$$\widehat{\gamma}_{SP}(\lambda, k) = \beta \cdot \widehat{\gamma}_{SP}(\lambda - 1, k) + (1 - \beta) \cdot \min[\widehat{\gamma}_I(\lambda, k), 30] \quad (23)$$

The term  $\widehat{\gamma}_{SP}(\lambda, k)$  is the approximate of smoothing of  $\gamma$ ,  $\beta$  is the constant of smoothing, and  $\widehat{\gamma}_I(\lambda, k)$  is shown in Eq. 22. The *min* function is used to limiting the value of  $\widehat{\gamma}_I(\lambda, k)$  to the maximum of 15 dB. If  $\beta = 0$ , then the value of  $\widehat{\gamma}_{SP}(\lambda, k) = \widehat{\gamma}_I(\lambda, k)$ . The values of estimation of  $\widehat{\xi}(\lambda, k)$  and  $\widehat{\gamma}_{SP}(\lambda, k)$  are exploited to estimate the gain function in Eq. 19. The signal transfer function is given in Eq. 19 is based on the instantaneous or current values of  $\gamma$  and  $\xi$ . The  $\gamma$  and  $\xi$  values may vary from one frame to another frame due to signal variation with respect to time. Under this condition, it is very difficult to evaluate those values with high performance and quality. In addition to this, we cannot calculate the value of  $\xi$ , since we do not have access directly to the spectrum of clean or original speech signal. Therefore, we use the past values or approximates of clean speech signal spectra to evaluate  $\xi$ . Given that the  $\gamma$  and  $\xi$  can be approximated either using Eq. 22 or Eq. 23, we shall estimate using both possibilities. If we consider both equations, then we will be having two transfer functions. First one is  $\widehat{H}_{SP_I}(\widehat{\xi}_i, \widehat{\gamma}_i)$ , which is mainly based on the instantaneous measurement values of  $\gamma$  and  $\xi$  shown in Eqs. 22 and 20 respectively. The term  $\widehat{H}_{SP}(\widehat{\xi}, \widehat{\gamma}_{SP})$  is the second transfer function which is based on the long term mean measurement values of  $\gamma$  and  $\xi$  shown in Eqs. 23 and 21, respectively. In summary, the proposed SP technique comprised of the following procedural stages:

- At stage 1: The magnitude spectrum of corrupted speech signal  $a_I(\lambda, k)$  is computed using fast Fourier transform (FFT) at frame  $\lambda$ .
- At stage 2: In [34], the authors have used the optimal smoothing and minimum statistics (OSMS) [20] technique for estimation of noise. This particular OSMS technique updates the estimated value of noise based on tracking statistics of local minimum of corrupted speech spectrum. Therefore, the time required for the adaptation for estimation of noise is equal to adaptation time required for the local minima. The method used (OSMS) in [34] is worked better for noise conditions where the power of noise is slowly vary with respect to time, but for sudden increment in noise levels, the OSMS technique requires more adaptation time by 1.5 seconds. But the method which we employed in our work requires only 0.52 seconds. Therefore, this noise approximation technique (computation time required=0.52 seconds) is used for updating the noise signal power spectrum  $[\widehat{a}_L(\lambda, k)]^2$ .
- At stage 3: Calculate  $\widehat{\gamma}_{SP}(\lambda, k)$  as per Eqs. 22 and 23.
- At stage 4: Use  $\widehat{\gamma}_{SP}(\lambda, k)$  to calculate  $\widehat{\xi}_{SP}(\lambda, k)$  as per Eq. 21. Flooring the value of  $\widehat{\xi}_{SP}(\lambda, k)$  to  $\xi_{\min}$  for the values of  $\widehat{\xi}_{SP}(\lambda, k)$  lesser than  $\xi_{\min}$ .
- At stage 5: Approximate the gain function  $\widehat{H}_{SP}(\widehat{\xi}, \widehat{\gamma}_{SP})$  using Eq. 19 and limit it to one.
- At stage 6: Get the processed or enhanced signal magnitude spectra by:  $\widehat{a}_J(\lambda, k) = \widehat{H}_{SP}(\widehat{\xi}, \widehat{\gamma}_{SP}) \cdot a_I(\lambda, k)$
- At stage 7: Calculate the inverse FFT of  $\widehat{a}_J(\lambda, k) \cdot e^{j\theta_I(\lambda, k)}$ , the term  $\theta_I(\lambda, k)$  is called the phase or angle of noisy speech signal to get the enhanced or processed

speech data.

The technique uses the transfer function  $\widehat{H}_{SP}(\widehat{\xi}, \widehat{\gamma}_{SP})$  is mainly based on instantaneous measurement values of  $\gamma$  and  $\xi$  can be developed by setting the parameters  $\beta = 0$  in Eq. 23 and  $\alpha = 1$  in Eq. 21. Therefore the instantaneous SP algorithm can be denoted as  $SP_I$  algorithm. The proposed SP technique is subjected to frames of 20 ms duration, and Hanning window is used for windowing the frames. The 50 % of overlapping rate is considered, and overlap-and-add method is used for reconstructing the processed speech signal. The constants for smoothing used in Eqs. 21 and 23 are set at  $\alpha = 0.98$  and  $\beta = 0.6$ . These values are not fixed; based on the experimental setup and performance measures, we change the values of  $\alpha$  and  $\beta$ . For the  $SP_I$  technique, the values of  $\alpha$  and  $\beta$  are kept to 1 and 0, respectively.

## 2.4 Performance Evaluation of Proposed and Existing Algorithms

The two types of performance measurement parameters are exploited in the present work to assess the performances of existing and proposed methods. The mean square error (MSE) is the first measure that we considered which is defined as the difference amidst spectrum magnitude of enhanced and clean speech signals. The second measures are objectives measure, namely the log likelihood ratio (LLR) and the perceptual evaluation of speech quality (PESQ).

### 2.4.1 Assessment Using MSE

The traditional formula for MSE is given by

$$MSE = \frac{1}{M \cdot N} \sum_{\lambda=0}^{M-1} \sum_{K=0}^{N-1} (a_J(\lambda, k) - \widehat{a}_J(\lambda, k))^2 \quad (24)$$

The terms  $a_J(\lambda, k)$  and  $\widehat{a}_J(\lambda, k)$  are the magnitude spectrum of original or clean speech data and enhanced processed speech data, respectively, at each frame  $\lambda$  and frequency bin  $k$ .  $M$  and  $N$  are designated as total number of frames and bins in a sentence. The MSE is calculated for proposed algorithm and compared with SS-VAD. According to Eq. 23, we have conducted experiments by varying the value of  $\beta$  from 0 to 1 and another smoothing constant  $\alpha$  is set at 0.98. In addition to this, we have considered the assessment of  $SP_I$  technique by setting  $\beta = 0$  and  $\alpha = 1$ . For the conduction and comparison of experiments with proposed method, we have developed SS-VAD algorithm and we have introduced another spectral subtractive algorithm called smoothed spectral subtraction (SSS) algorithm by replacing  $\widehat{\gamma}(k)$  in Eq. 5 with its smoothed variety given in Eq. 23. For the conduction and evaluation of experiments for the proposed and existing algorithms, the NOIZEUS speech corpora [10] and continuous Kannada speech sentences were used. The conducted experimental results are shown in Tables 1 and 2 in terms of MSE for NOIZEUS and Kannada continuous speech database, respectively. From both the tables, we observe that the proposed SP technique has given lesser values of MSE compared to SSS and SS-VAD algorithms at lower SNR levels (0 and 5 dB) at  $\beta = 0.98$ , which clearly shows that the smoothing

**Table 1** The MSE results obtained by SP,  $SP_I$ , SS-VAD and SSS algorithms for NOIZEUS database

Method	SNR (dB)	$\beta=0$	$\beta=0.15$	$\beta=0.35$	$\beta=0.55$	$\beta=0.75$	$\beta=0.98$	$SP_I (\alpha=1, \beta=0)$
SP	0	2.73	2.67	2.69	2.68	2.39	1.95	4.50
SSS	0	4.23	4.13	4.11	3.98	3.77	3.75	4.30
SS-VAD	0	4.16	4.24	4.21	4.25	4.28	4.19	4.29
SP	5	1.49	1.50	1.49	1.53	1.34	1.01	2.77
SSS	5	1.22	1.30	1.26	1.25	1.20	1.33	1.30
SS-VAD	5	1.32	1.31	1.32	1.36	1.30	1.33	1.30
SP	10	0.79	0.85	0.86	0.87	0.78	0.60	1.59
SSS	10	0.42	0.40	0.41	0.38	0.39	0.48	0.40
SS-VAD	10	0.43	0.40	0.42	0.40	0.40	0.41	0.40

**Table 2** The MSE results obtained by SP,  $SP_I$ , SS-VAD and SSS algorithms for Kannada speech database

Method	SNR (dB)	$\beta=0$	$\beta=0.15$	$\beta=0.35$	$\beta=0.55$	$\beta=0.75$	$\beta=0.98$	$SP_I (\alpha=1, \beta=0)$
SP	0	2.70	2.61	2.63	2.61	2.39	1.94	4.49
SSS	0	4.22	4.12	4.12	3.99	3.79	3.79	4.33
SS-VAD	0	4.18	4.28	4.27	4.29	4.24	4.20	4.33
SP	5	1.45	1.50	1.48	1.53	1.35	0.99	2.70
SSS	5	1.21	1.32	1.27	1.26	1.21	1.29	1.31
SS-VAD	5	1.33	1.31	1.33	1.35	1.33	1.35	1.33
SP	10	0.76	0.82	0.85	0.82	0.79	0.61	1.60
SSS	10	0.43	0.44	0.44	0.39	0.39	0.42	0.42
SS-VAD	10	0.43	0.43	0.42	0.41	0.48	0.42	0.41

of  $\hat{\gamma}(\lambda, k)$  helped to minimize the MSE value. Higher MSE values were obtained from the proposed algorithm when the parameter  $\xi$  and  $\gamma$  were not smoothed. This reveals that the usage of present and past spectral characteristics information of signal could be more efficient than using only instantaneous values. In summary, the gain function derived in Eq. 19 remains very efficient at low SNR levels ranging from 0–5 dB, where the function of gain of the SS algorithm becomes less accurate at low SNR (0 dB).

#### 2.4.2 Evaluation of Speech Quality

The proposed SP algorithm is assessed using LLR [11] and PESQ measures [12] to check the quality of speech. The NOIZEUS speech database [10] and continuous Kannada speech database are considered for the conduction of experiments, which are sampled at a sampling frequency of 8 kHz and degraded by street noise, babble noise, car noise, train noise etc, taken from AURORA database [8] at different SNR (0, 5 and 10 dB). The objective measures PESQ and LLR got the correlation coefficient values of  $\rho = 0.67$  and  $\rho = 0.61$ , respectively, to evaluate the quality of speech [9, 11, 12]. Here also we try to compare the performance and reliability of proposed

**Table 3** The objective assessment (PESQ values) [12] and comparison of the proposed SP technique against SS-VAD, SSS, MMSE-SP, MMSE-SPZC and MAP estimators for NOIZEUS database

Method	Type of Noise	SNR=0 dB		SNR=5 dB		SNR=10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
SP	Car	1.85	1.80	2.20	2.15	2.58	2.57
$SP_I$		1.66		1.75		2.09	
SS-VAD		1.68	1.69	2.00	2.01	2.30	2.30
SSS		1.71	1.68	2.01	1.99	2.33	2.33
MMSE-SP		1.81	1.82	2.08	2.09	2.38	2.37
MMSE-SPZC		1.86	1.85	2.21	2.22	2.65	2.66
MAP	Babble	1.89	1.90	2.19	2.15	2.48	2.45
SP		1.83	1.79	2.17	2.10	2.54	2.55
$SP_I$		1.66		1.85		2.16	
SS-VAD		1.72	1.73	2.03	2.02	2.36	2.37
SSS		1.77	1.77	2.07	2.07	2.38	2.37
MMSE-SP		1.72	1.73	2.09	2.10	2.36	2.37
MMSE-SPZC	1.75	1.75	2.12	2.11	2.56	2.55	
MAP	Street	1.79	1.80	2.05	2.11	2.40	2.43
SP		1.76	1.70	2.17	2.01	2.52	2.52
$SP_I$		1.43		1.80		2.10	
SS-VAD		1.69	1.70	2.01	2.02	2.35	2.35
SSS		1.71	1.69	2.01	1.99	2.37	2.27
MMSE-SP		1.69	1.70	2.08	1.99	2.35	2.31
MMSE-SPZC	1.80	1.80	2.19	2.20	2.56	2.56	
MAP		1.74	1.77	2.12	2.09	2.39	2.40

technique against the traditional SS-VAD, SSS algorithms and magnitude squared spectrum (MSS) estimators [31]. The MSS estimators which are implemented in [31] are MMSE-spectrum power (MMSE-SP), MMSE-spectrum power estimator based on zero crossing (MMSE-SPZC) and maximum *a posteriori* (MAP). The implemented MSS estimators were modelled using Gaussian statistical model, and these algorithms cannot be directly compared with the proposed SP algorithm, because as it depends on various assumptions and principles. All approaches or methods were tested using two different values of  $\beta$  (0.6 and 0.98) and at  $\alpha = 0.98$ . Tables 3, 4, 5 and 6 show the objective measures in terms of PESQ and LLR for NOIZEUS (Table 3 and 5) and continuous Kannada speech database (Tables 4 and 6) for proposed technique against SS-VAD, SSS, MMSE-SP, MMSE-SPZC and MAP estimators, respectively.

The better performance assessment in terms of PESQ is represented by higher values and the lower values indicate that the high performance in case of LLR measure. The reliability of MMSE-SPZC and sometimes MAP algorithms was significantly better than that of the SP technique in most degraded situations, except in the case of babble at 0 and 5 dB SNR (shown in Tables 3 and 4). The process of smoothing of  $\gamma$  in MMSE-SP algorithm gives lesser values in terms of PESQ under all condi-

**Table 4** The objective assessment (PESQ values) [12] and comparison of the proposed SP technique against SS-VAD, SSS, MMSE-SP, MMSE-SPZC and MAP estimators for Kannada speech database

Method	Type of Noise	SNR=0 dB		SNR=5 dB		SNR=10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
SP	White	1.96	1.86	2.26	2.21	2.57	2.52
$SP_I$		1.45		1.74		2.12	
SS-VAD		1.61	1.65	1.93	1.95	2.29	2.27
SSS		1.66	1.66	1.94	1.92	2.33	2.30
MMSE-SP		1.76	1.76	2.01	2.09	2.30	2.31
MMSE-SPZC		1.99	2.00	2.32	2.35	2.69	2.70
MAP	Car	1.85	1.85	2.36	2.37	2.45	2.45
SP		1.84	1.81	2.22	2.20	2.60	2.56
$SP_I$		1.65		1.77		2.09	
SS-VAD		1.67	1.67	1.99	2.00	2.29	2.28
SSS		1.69	1.67	2.00	1.98	2.32	2.31
MMSE-SP		1.80	1.81	2.01	2.08	2.32	2.34
MMSE-SPZC	Babble	1.87	1.86	2.25	2.20	2.68	2.69
MAP		1.86	1.89	2.18	2.16	2.50	2.49
SP		1.89	1.87	2.19	2.20	2.59	2.56
$SP_I$		1.66		1.84		2.17	
SS-VAD		1.70	1.72	2.01	2.05	2.35	2.36
SSS		1.76	1.76	2.08	2.09	2.36	2.35
MMSE-SP	Street	1.70	1.72	2.10	2.11	2.29	2.34
MMSE-SPZC		1.76	1.74	2.11	2.11	2.54	2.54
MAP		1.78	1.80	2.04	2.10	2.38	2.40
SP		1.76	1.70	2.18	2.14	2.53	2.51
$SP_I$		1.42		1.79		2.11	
SS-VAD		1.70	1.71	2.00	2.07	2.36	2.37
SSS	1.72	1.71	2.07	2.00	2.39	2.27	
MMSE-SP	Street	1.68	1.69	2.07	2.00	2.36	2.33
MMSE-SPZC		1.87	1.87	2.20	2.28	2.60	2.60
MAP		1.78	1.79	2.15	2.10	2.40	2.44

tions. The SP algorithm has an advantage compared to MSS estimators, because of its low complexity computation, less memory and time for execution as it needs a few multiplication and addition operations (Eq. 19). The MSS estimators, on the other hand, need an implementation of Bessel functions. The proposed SP technique implemented at  $\beta = 0.6$  worked well and constantly better than the SS-VAD and SSS under all degraded environments. The SP technique has given unsatisfactory results when it is implemented with  $\beta = 0.98$ , specifically at the lower SNR levels. This implies that the SP technique is much sensitive to  $\beta$  value used for approximating and updating the parameter  $\hat{\gamma}(\lambda, k)$ . The  $\beta$  value is equal to 0.6 gives approximately equal weight to the

**Table 5** The objective assessment (LLR values) [12] and comparison of the proposed SP technique against SS-VAD, SSS, MMSE-SP, MMSE-SPZC and MAP estimators for NOIZEUS database

Method	Type of Noise	SNR=0 dB		SNR=5 dB		SNR=10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
SP	White	1.55	1.60	1.27	1.30	1.07	1.08
$SP_I$		1.72		1.43		1.19	
SS-VAD		1.73	1.73	1.48	1.48	1.22	1.22
SSS		1.72	1.69	1.44	1.40	1.19	1.18
MMSE-SP		1.71	1.70	1.45	1.41	1.20	1.19
MMSE-SPZC		1.56	1.57	1.26	1.30	1.05	1.08
MAP	Car	1.84	1.86	1.49	1.50	1.33	1.28
SP		0.97	1.01	0.79	0.82	0.65	0.64
$SP_I$		1.23		1.06		0.88	
SS-VAD		1.01	1.01	0.78	0.78	0.59	0.58
SSS		0.99	0.99	0.77	0.76	0.57	0.57
MMSE-SP		1.08	1.08	0.78	0.79	0.59	0.60
MMSE-SPZC	Babble	1.01	1.01	0.78	0.78	0.62	0.62
MAP		1.11	1.10	0.93	0.88	0.77	0.78
SP		0.92	0.99	0.69	0.81	0.68	0.69
$SP_I$		1.17		1.02		0.82	
SS-VAD		0.95	0.95	0.76	0.77	0.55	0.55
SSS		0.93	0.94	0.74	0.73	0.53	0.53
MMSE-SP	Street	0.93	0.95	0.75	0.77	0.55	0.56
MMSE-SPZC		1.15	1.14	0.90	0.90	0.67	0.67
MAP		1.23	1.24	1.01	0.99	0.84	0.84
SP		1.03	1.11	0.83	0.87	0.69	0.70
$SP_I$		1.23		1.05		0.84	
SS-VAD		1.01	1.01	0.81	0.80	0.63	0.63
SSS	1.00	0.99	0.78	0.80	0.59	0.61	
MMSE-SP	Street	1.11	1.11	0.80	0.81	0.67	0.67
MMSE-SPZC		1.00	1.01	0.87	0.88	0.68	0.67
MAP		1.19	1.14	1.05	0.99	0.80	0.90

exploitation of past and spectral characteristic information when evaluating  $\hat{\gamma}(\lambda, k)$ . In comparison, the performance of the SS-VAD technique was not much affected when  $\hat{\gamma}(\lambda, k)$  parameter was nicely smoothed. The  $SP_I$  technique based on the instantaneous spectral characteristics measurements of  $\gamma$  and  $\xi$ , given unsatisfactory values under all degraded conditions. Of course it is true because, the instantaneous values of  $\gamma$  and  $\xi$  may change dramatically from one frame to another frame which causes high level of musical noise [4] resulting from sudden variations of the function of gain. This reveals that the smoothing parameters of  $\gamma$  and  $\xi$  are much indeed to get high quality of speech free of musical noise or musical tones. This reliability was much similar when

**Table 6** The objective assessment (LLR values) [12] and comparison of the proposed SP technique against SS-VAD, SSS, MMSE-SP, MMSE-SPZC and MAP estimators for Kannada speech database

Method	Type of Noise	SNR=0 dB		SNR=5 dB		SNR=10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
SP	White	1.54	1.59	1.26	1.31	1.08	1.08
$SP_I$		1.73		1.43		1.19	
SS-VAD		1.74	1.73	1.49	1.50	1.23	1.24
SSS		1.73	1.70	1.43	1.43	1.20	1.20
MMSE-SP		1.72	1.72	1.48	1.43	1.23	1.20
MMSE-SPZC		1.52	1.55	1.25	1.30	1.0	1.09
MAP	Car	1.83	1.85	1.50	1.49	1.32	1.28
SP		0.95	1.00	0.75	0.83	0.66	0.68
$SP_I$		1.21		1.05		0.89	
SS-VAD		1.02	1.03	0.80	0.81	0.60	0.59
SSS		1.00	1.01	0.80	0.81	0.59	0.58
MMSE-SP		1.09	1.08	0.79	0.80	0.60	0.62
MMSE-SPZC	1.00	1.00	0.80	0.80	0.63	0.63	
MAP	Babble	1.10	1.09	0.93	0.88	0.77	0.79
SP		0.89	1.09	0.71	0.92	0.65	0.68
$SP_I$		0.99		0.99		0.83	
SS-VAD		0.98	1.01	1.02	1.01	1.03	1.05
SSS		1.01	1.01	1.03	1.10	1.12	1.11
MMSE-SP		0.93	0.95	0.75	0.77	0.55	0.56
MMSE-SPZC	1.14	1.11	0.90	0.90	0.67	0.67	
MAP	Street	1.22	1.21	0.99	0.99	0.84	0.83
SP		1.01	1.09	0.83	0.84	0.70	0.71
$SP_I$		1.22		1.06		0.89	
SS-VAD		1.02	1.01	0.83	0.81	0.65	0.69
SSS		1.01	1.01	0.85	0.80	0.64	0.62
MMSE-SP		1.12	1.12	0.82	0.81	0.68	0.68
MMSE-SPZC	1.01	1.02	0.86	0.87	0.69	0.69	
MAP		1.20	1.13	1.06	1.00	0.81	0.93

the proposed and existing techniques were assessed using the LLR measure shown in Tables 5 and 6 for NOIZEUS and continuous Kannada speech corpora, respectively.

In summary, it is well clear that the proposed SP technique has given lesser residual noise compared to spectral subtractive techniques. Informal-listening-tests were also conducted for the experimental purpose (to keep simplicity, those tests are not mentioned in the paper) which reveals that the enhanced speech data obtained from the proposed SP technique comprised of a smoother background with no audibility of various types of background noises and musical noise at different SNR levels. As per the implementation of proposed algorithm, we trust that the SP technique does



**Table 7** Numerical complexity comparison of algorithms used in this work for the speech data degraded by babble noise at different SNR levels.

Computation Parameters	Algorithms	SNR = 0 dB	SNR = 5 dB	SNR = 10 dB
Memory consumption in bytes	SS-VAD	758625	742568	712548
	SSS	745895	739945	711224
	MMSE-SP	669344	712704	701288
	MMSE-SPZC	655360	624288	623440
	MAP	655360	647168	635216
	$SP_I$	633245	632147	623546
	SP	<b>601234</b>	<b>592146</b>	<b>587946</b>
Elapsed time in seconds	SS-VAD	0.3513	0.3412	0.3262
	SSS	0.3555	0.3411	0.3125
	MMSE-SP	0.0379	0.0301	0.0630
	MMSE-SPZC	0.0475	0.0486	0.0316
	MAP	0.0486	0.0923	0.0680
	$SP_I$	0.0345	0.0300	0.0301
	SP	<b>0.0301</b>	<b>0.0297</b>	<b>0.0291</b>

Bold values indicate the best performance

not have audibility of background and musical noises because it takes some characteristics of MSS estimators. In contrast to this, SS algorithms have given poor results in suppression of musical noise. From the conducted and investigated experimental results, it can be inferred that the proposed SP technique has reduced various types of background noises in degraded speech data (NOIZEUS and continuous Kannada speech databases) at different SNR levels. Therefore, the proposed noise reduction technique could be introduced at the front end of continuous Kannada ASR system.

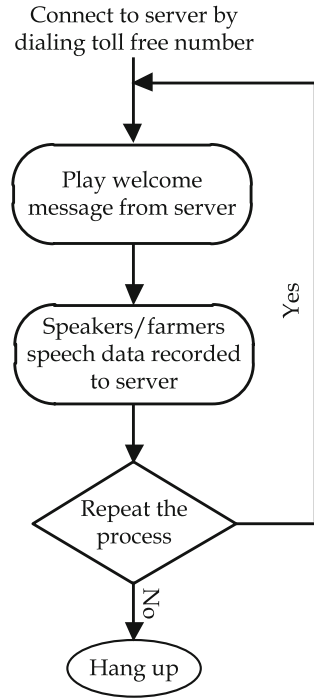
### 2.4.3 Numerical Complexity Comparison

In this work, memory consumption and average execution time are used as numerical complexity comparison metrics for existing and proposed speech enhancement techniques. The considered algorithm(s) is executed around ten times and noted the memory consumption values in bytes. The mean of the ten values is taken into consideration for the calculation of final memory consumption. Similarly, the time in seconds required to execute the algorithm is computed. From Table 7, it is observed that the SP speech enhancement technique consumes less memory and execution time as compared to the other methods.

## 3 Robust Continuous Kannada ASR System by Proposed Noise Elimination Algorithm

For the paper completeness, the earlier end-to-end continuous Kannada speech recognition system [17] modules are explained in this section briefly. For the creation of

**Fig. 3** The structure of call flow for the speech data collection



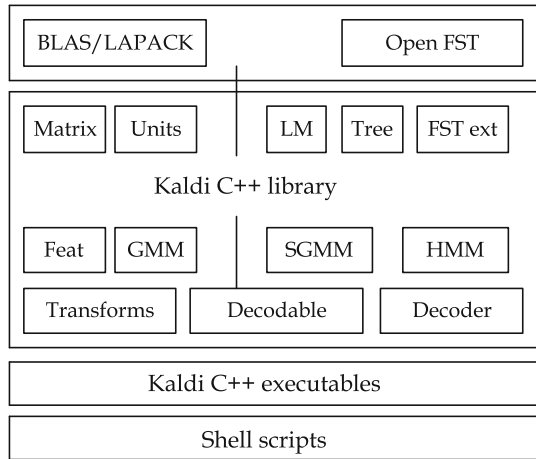
robust ASR system, the speech data collection plays an important role. Each continuous speech sentence will be having various pronunciations. Keeping this in context, the authors have collected the continuous Kannada speech data from 2400 speakers/farmers under noisy conditions. The interactive voice response system (IVRS) structure of call flow is implemented for the collection of speech data. The decision making processes are considered by the call flow shown in Fig. 3. The farmers/speaker need to dial the toll free number to have connect to the server. In the process of call getting connected to the server, the server will play-out the pre-recorded prompt that, “Welcome to continuous Kannada speech data collection centre, please tell the names of continuous Kannada speech sentences after the beep sound”. The server will prompt a total of 40 continuous speech sentences in each session for 2.5 minutes. Followed by the beep sound from the server, the speaker/farmer needs to repeat the sentences. Once all 40 sentences are completed in the call flow, the server will instruct the speaker to repeat the process. If the speaker is not satisfied with his/her speech recordings, then he/she can repeat the process by saying “Yes”. If he/she says “Yes”, then it will go to first step; otherwise the call will hang up. The continuous Kannada speech sentences used for speech database collection are shown in Table 8.

Nowadays, Kaldi [24] is the most widely used toolkit for the creation of ASR models. It has various features and recent modelling techniques compared to other speech recognition toolkits. The block diagram of Kaldi is shown in Fig. 4. It consists of three important modules, namely: the front end, the linguistic and the decoder. The front end receives the transcribed and validated speech data to extract the speech features

**Table 8** Continuous speech sentences used for data collection

English version of continuous Kannada speech sentences	Kannada Version
Adda godeya mele deepa itta haage	ಅಡ್ಡ ಗೋಡೆಯ ಮೇಲೆ ದೀಪ ಇಟ್ಟ ಹಾಗೆ
Athi aase gathi gedu	ಅತಿ ಆನೆ ಗತಿ ಗೇಡು
Bandaddella barali govindana daye irali	ಬಂದಡ್ಡಲ್ಲಾ ಬರಲಿ ಗೋವಿಂದನ ದಯೆ ಇರಲಿ
Akki mele aase nentara mele preeti	ಅಕ್ಕಿ ಮೇಲೆ ಆನೆ ನೆಂಟರ ಮೇಲೆ ಪ್ರೀತಿ
Beline eddu hola meytante	ಬೇಲಿನೆ ಎದ್ದು ಹೊಲ ಮೇಯಿತಂತೆ
Haavu saayalilla kolu muriyalilla	ಹಾವು ಸಾಯಲಿಲ್ಲ ಕೋಲು ಮುರಿಯಲಿಲ್ಲ
Veda sulladaru gaade sullagadu	ವೇದ ಸುಳ್ಳಾದರು ಗಾಡೆ ಸುಳ್ಳಾಗದು
Adikege hoda maana aane kottaru baaradu	ಅಡಿಕೆಗೆ ಹೋದ ಮಾನ ಆನೆ ಕೊಟ್ಟರು ಬಾರದು
Haagalakaayige bevinakaayi saakshi	ಹಾಗಲಕಾಯಿಗೆ ಬೇವಿನಕಾಯಿ ಸಾಕ್ಷಿ
Enne bandaga kannu mucchhikondante	ಎಣ್ಣೆ ಬಂದಾಗ ಕಣ್ಣು ಮುಚ್ಚಿಕೊಂಡಂತೆ
Hiriyakkana chaali mane mandigella	ಹಿರಿಯಕ್ಕನ ಚಾಳಿ ಮನೆ ಮಂದಿಗಲ್ಲ
Hoovininda naaru swarga seritu	ಹೊವಿನಿಂದ ನಾರು ಸ್ವರ್ಗ ಸೇರಿತು
Bekkige chellata ilige praana sankata	ಬೆಕ್ಕಿಗೆ ಚೆಲ್ಲಾಟ ಇಲಿಗೆ ಪ್ರಾಣ ಸಂಕಟ
Hana andre henavu baayi biduttade	ಹಣ ಅಂದ್ರೆ ಹೆಣವು ಬಾಯಿ ಬಿಡುತ್ತದೆ
Maatu belli mouna bangara	ಮಾತು ಬೆಳ್ಳಿ ಮೌನ ಬಂಗಾರ
Tumbida koda tulukuvudilla	ತುಂಬಿದ ಕೊಡ ತುಳುಕುವುದಿಲ್ಲ
Hani hani seridare halla tene tene seridare balla	ಹನಿ ಹನಿ ಸೇರಿದರೆ ಹಳ್ಳ ತನ ತನ ಸೇರಿದರೆ ಬಳ್ಳು
Attegondu kaal sosegondu kaala	ಅತ್ತೆಗೊಂದು ಕಾಲ ಸೊಸೆಗೊಂದು ಕಾಲ
Chinte illadavanige santelu nidde	ಚಿಂತೆ ಇಲ್ಲದವನಿಗೆ ಸಂತೇಲೂ ನಿಡ್ಡೆ
Ollada gandani mosaralli kallu	ಒಲ್ಲದ ಗಂಡನಿಗೆ ಮೊಸರಲ್ಲಿ ಕಲ್ಲು
Kai keasaradare baayi mosaru	ಕೈ ಕೆಸರಾದರೆ ಬಾಯಿ ಮೊಸರು
Huttu guna suttaru hogalla	ಹುಟ್ಟು ಗುಣ ಸುಟ್ಟರು ಹೋಗಲ್ಲ
Hannele uduruvaaga chigurele nagutittu	ಹಣ್ಣೆಲೆ ಉದುರುವಾಗ ಚಿಗುರಲೆ ನಗುತ್ತಿತ್ತು
Vinasha kale vipareeta buddi	ವಿನಾಶ ಕಾಲಕ್ಕೆ ವಿಪರೀತ ಬುದ್ಧಿ
Yarado duddu yellamma jaatre	ಯಾರದೋ ದುಡ್ಡು ಎಲ್ಲಮ್ಮನ ಜಾತ್ರೆ
Hittala gida maddalla	ಹಿತ್ತಲ ಗಿಡ ಮದ್ದಲ್ಲ
Haaloorige ulidavane gouda	ಹಾಳೂರಿಗೆ ಉಳಿದವನೇ ಗೌಡ
Holeyuvudella chinnavalla	ಹೊಳೆಯುವುದಲ್ಲಾ ಚಿನ್ನವಲ್ಲ
Raatri kanda baavili hagalu biddange	ರಾತ್ರಿ ಕಂದ ಬಾವಿಲಿ ಹಗಲು ಬಿದ್ದಂಗ
Uppu tinda mele neeru kudiya beku	ಉಪ್ಪು ತಿಂದ ಮೇಲೆ ನೀರು ಕುಡಿಯಲೇ ಬೇಕು
Indina solu naalina geluvu	ಇಂದಿನ ಸೋಲು ನಾಳಿನ ಗೆಲುವು
Gidavagi baggadaru maravagi baggitte	ಗಿಡವಾಗಿ ಬಗ್ಗಿದ್ದು ಮರವಾಗಿ ಬಗ್ಗಿತೇ
Manasiddare maarga	ಮನಸಿದ್ದರೆ ಮಾರ್ಗ
Aapatigadavane nijavada geleya	ಆಪತ್ತಿಗಾದವನೇ ನಿಜವಾದ ಗೆಳೆಯ
Namma devara satya namage gottu	ನಮ್ಮ ದೇವರ ಸತ್ಯ ನಮಗೆ ಗೊತ್ತು
Badavana sattu davadege moola	ಬಡವನ ಸಿಟ್ಟು ದವಡೆಗೆ ಮೂಲ
Gaaada mele bare eledante	ಗಾಯದ ಮೇಲೆ ಬರೆ ಎಳೆದಂತೆ
Aaseye dukhakke moola	ಆಸೆಯೇ ದುಃಖಕ್ಕೆ ಮೂಲ
Atiyadare amrutavu visha	ಅತಿಯಾದರೆ ಅಮೃತವೂ ವಿಷ

Fig. 4 Block diagram of Kaldi



using Mel frequency cepstral coefficient (MFCC) technique. The linguistic consists of acoustic features and language resources such as lexicon, phoneme set, silence phones and non-silence phones. The decoder mainly uses the acoustic features of speech, and testing is done in this module. To get the transcriptions of each continuous speech sentence, we need a particular language phoneme set and lexicon. The lexicon will be consisting of two parts, word-level transcription (left-part) and phoneme-level transcription (right-part). The left part of the dictionary/lexicon is obtained using symbols of IT3:UTF-8 transliteration tool. The right part of the lexicon is built using the Indian language speech sound label (ILSL12) set shown in Table 9.

Using these phoneme sets, we can easily create the lexicon for continuous Kannada speech data shown in Table 10. Some of the tags are used for the representation of background noises which are added in the speech data since it is collected under an uncontrolled environment. The different tags and their meanings are described in Table 11. The content of the speech file needs to be converted to its equivalent text format, and that process is called transcription. Transcriptions of some of the continuous Kannada speech sentences are shown in Fig. 5. All transcriptions are done using a transcribing tool which is also shown in Fig. 5. The continuous speech features are extracted using the acoustic MFCC technique. The following are the parameters are used for the extraction of speech features:

- 39-dimensional feature vector.
- Sampling frequency of 8 kHz.
- Hanning window for windowing procedure.
- The FFT for transforming the speech signal from time to frequency domain.
- Mel filter bank for linear to logarithmic.
- Discrete cosine transform for transforming the speech signal from frequency to time domain.

The 67,200 and 16,800 of transcribed degraded speech sentences were used for Kaldi system training and decoding, respectively [17]. The following are the parameters were used for the creation of language models (LMs) and acoustic models (AMs):

**Table 9** The symbols used from transliteration tool and ISL12 to create lexicon

Label set using IT3: UTF-8				Corresponding Kannada phonemes				Label set using ILSL12				Corresponding Kannada phonemes			
a	oo	t:h	ph	ಅ	ಓ	ಠ	ಫ	a	oo	txh	ph	ಅ	ಓ	ಠ	ಫ
aa	au	d	b	ಆ	ಔ	ಡ	ಬ	aa	au	dx	b	ಆ	ಔ	ಡ	ಬ
i	k	d:h	bh	ಇ	ಫ	ಢ	ಭ	i	k	dxh	bh	ಇ	ಫ	ಢ	ಭ
ii	kh	nd~	m	ಈ	ಖ	ಣ	ಮ	ii	kh	nx	m	ಈ	ಖ	ಣ	ಮ
u	g	t	y	ಉ	ಗ	ತ	ಯ	u	g	t	y	ಉ	ಗ	ತ	ಯ
uu	gh	th	r	ಊ	ಘ	ಠ	ರ	uu	gh	th	r	ಊ	ಘ	ಠ	ರ
e	c	d	l	ಎ	ಚ	ದ	ಲ	e	c	d	l	ಎ	ಚ	ದ	ಲ
ee	ch	dh	v	ಏ	ಛ	ಢ	ವ	ee	ch	dh	w	ಏ	ಛ	ಢ	ವ
								ai	j	n	sh	ಐ	ಜ	ನ	ಶ
								o	tx	p	s	ಒ	ಟ	ಪ	ಸ

**Table 10** Lexicon for some of the continuous Kannada speech sentences

Label set using IT3: UTF-8	Label set using ILSL12
veida sul:l:l:aadaru gaade sul:l:l:aagadu	w ee d a_s u llx a d r u_g aa d e_s u llx aa g d u
ad:i:kege hooda maana aane kot:t:aruu baaradu	a dx i k e e_h oo d a_maa n a_aa n e_k o t t x a r u_b a a r a d u
kai kesaraadare baayi mosaru	k ai_k e s a r a a d a r e_b a a y i_m o s a r u
maatu bel:l:i mauna ban:gaara	m aa t u_b e l l x i_m a u n a_b a n g g a a r a
ad:d:a good:eya meile diipa it:t:a haage	a d dx_a_g oo dx e y a_m e i l e_d i i p a_i t t x a_h a a g e
akki meile aase nen:t:ara meile priiti	a k k i_m e i l e_a a s e_n e n t x a r a_m e i l e_p r i i t i

**Table 11** The different tags and their explanation

Tag	Description
<s>	Starting of the speech sentence
</s>	Ending of the speech sentence
<pau>	Pauses in amidst of speech utterances
<aah>	Sound generated when the mouth is opened
<hmm>	Sound generated when the mouth is closed
<hm>	Its just like saying “yes”
<hm-hm>	Its just like saying “no”
<laugh>	When the speaker laughs in amidst sentence
<horn>	Vehicle sound added in the speech sentences
<babble>	When multi speakers speech is added to intended speaker’s speech
<bn>	Noise produced in background
<n>	Vocal noise



MNKKKN01I1000 maatu\_bel:l:i\_mauna\_ban:gaara  
 MNKKKN01I1001 ad:d:a\_good:eya\_mele\_diipa\_it:t:a\_haage  
 MNKKKN01I1002 kai\_kesaraadare\_baayi\_mosaru  
 MNKKKN01I1003 ati\_aase\_gati\_geid:u  
 MNKKKN01I1004 beiline\_eddu\_hola\_meiyitan:te  
 MNKKKN01I1005 haal:uurige\_ul:idavane\_gawd:a  
 MNKKKN01I1006 haagalakaayige\_bevinakaayi\_saakshi  
 MNKKKN01I1007 haavu\_saayalilla\_koolu\_murililla  
 MNKKKN01I1008 hittala\_gid:a\_maddalla  
 MNKKKN01I1009 hut:t:u\_gund~a\_sut:t:taru\_hoogolla  
 MNKKKN01I1010 vinaasha\_kaalakke\_vipariita\_buddhi

Fig. 5 Transcribing tool and some of the continuous speech sentences transcription

- Lexicon: It can also be called as dictionary which acts as base for the speech recognition accuracy. It consists of both word level and phoneme level transcriptions of degraded speech data.
- Phones for silence: The words SIL and sil could be considered as silence phones.
- <s> and </s>: These indicates that the starting and ending of the speech sentences.
- Non silence phones or speech phonemes: Totally 168 speech sounds/phonemes were exploited for the creation of lexicon and transcription.
- Triphone 1 leaves: 500.
- Triphone 1 Gaussian mixtures: 2000.
- Maximum likelihood linear transform (MLLT) leaves: 500.
- MLLT Gaussian mixtures: 2000.
- Speaker adaptation technique (SAT) leaves: 500.
- SAT Gaussian leaves: 2000.
- Universal background model (UBM) Gaussian mixtures: 200.
- Subspace Gaussian mixture model (SGMM) leaves: 3000.
- SGMM Gaussian mixtures: 3000.
- Training and testing jobs: 3.
- Deep neural network (DNN) hidden layers: 2.

The continuous Kannada speech ASR models for Kannada language were obtained at following phoneme levels:

- Training and decoding at single phone level.
- Triphone 1: Delta+Delta-Delta training and decoding.
- Triphone 2: linear discriminant analysis (LDA) + MLLT training and testing.

**Table 12** The least WERs obtained using Kaldi for continuous Kannada speech corpora of previous work [17] and present work

Phoneme levels	WER as in [17]	WER of present work
Mono phone	7.66	7.06
Triphone1-600-2400	6.29	5.01
Triphone1-600-4800	5.63	5.04
Triphone1-600-9600	5.09	5.05
Triphone2-600-2400	7.00	5.68
Triphone2-600-4800	5.96	5.67
Triphone2-600-9600	5.80	5.66
Triphone3-600-2400	6.18	5.05
Triphone3-600-4800	5.69	5.02
Triphone3-600-9600	5.03	4.99
SGMM	4.65	4.45
Combined SGMM and MMI with Iteration1	4.60	4.44
Combined SGMM and MMI with Iteration2	4.70	4.48
Combined SGMM and MMI with Iteration3	4.60	4.40
Combined SGMM and MMI with Iteration4	4.65	4.41
Combined DNN and HMM	<b>4.10</b>	<b>2.91</b>
Combined DNN and SGMM with Iteration1	4.21	3.01
Combined DNN and SGMM with Iteration2	4.21	2.96
Combined DNN and SGMM with Iteration3	4.27	2.99
Combined DNN and SGMM with Iteration4	4.27	2.98
TDNN	–	<b>2.46</b>
Proposed speech enhancement algorithm + TDNN	–	<b>2.23</b>

Bold values indicate the best performance

- Triphone 3: LDA + MLLT + SAT training and testing.
- SGMM training and testing.
- Hybrid training and testing using DNN (DNN+HMM).
- DNN + SGMM with various iterations.
- SGMM + maximum mutual information (MMI) with various iterations.
- TDNN

Table 12 shows the WER obtained for various modelling techniques for continuous Kannada noisy and enhanced speech data. From Table, it can be observed that the combination of DNN and HMM has given the least WER of 4.10% which outperforms all other modelling techniques in our earlier work. Though the WER was less for the degraded continuous Kannada speech data, we believed that we can further minimize the WER by applying the proposed noise reduction algorithm and recent modelling technique called TDNN on corrupted continuous Kannada speech database. Upon experiments, the combination of DNN and HMM has given the least WER of 2.91% compared to other modelling techniques for enhanced speech data. Further the combination of proposed speech enhancement technique and TDNN has reduced the

WER to 2.23%. Therefore, it can be inferred that, the amalgamation of TDNN and speech enhancement technique has given significant improvement in WER of 1.87% compared to earlier DNN-HMM-based continuous Kannada SQS ASR models.

The block diagram of newly developed SQS is depicted in Fig. 6. It consists of three important modules, namely, asterisk server, IVRS call flow and ASR models.

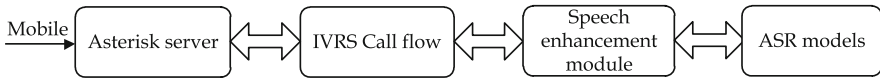
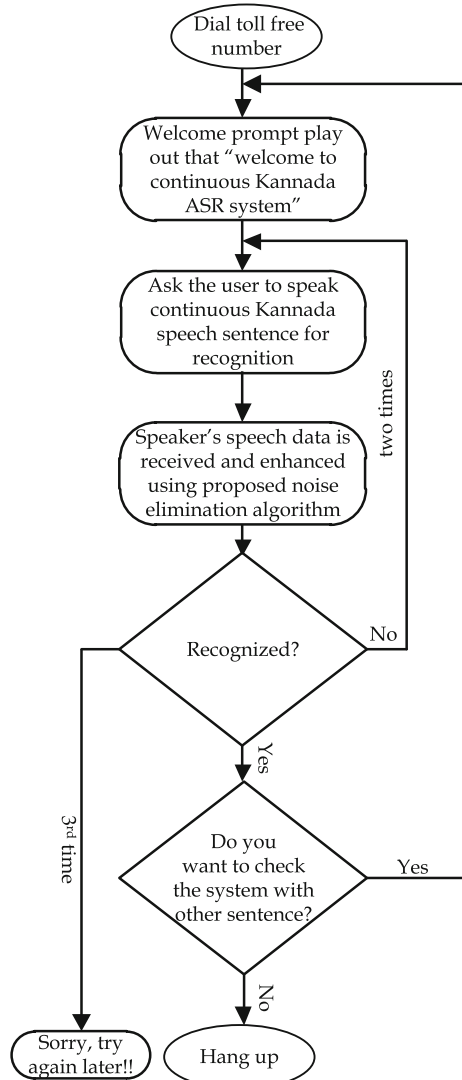


Fig. 6 Block diagram of newly developed SQS system

Fig. 7 Call flow structure of newly developed continuous Kannada SQS





**Table 13** Online assessment of developed SQS system by users

Total continuous Kannada speech sentences used for speech data collection	40
Total number of users tested the system	500
<i>Earlier Kannada continuous speech SQS</i>	
Recognitions at first time	430
Recognitions at second time	34
Recognitions at third time	15
Total speech recognitions	479
Accuracy of speech recognition in percentage	95.80%
<i>Current Kannada continuous speech SQS</i>	
Recognitions at first time	455
Recognitions at second time	26
Recognitions at third time	07
Total speech recognitions	488
Accuracy of speech recognition in percentage	97.60%

The asterisk server is used to connect the mobile calls to the server, and IVRS call flow structure gives the entire structure and decision making process of SQS. Finally, the last module gives the least WER continuous Kannada ASR models for enhanced speech data built by proposed noise elimination technique and Kaldi. The call flow structure of SQS is shown in Fig. 7. The user/farmer needs to dial the toll free number. Once the call gets connected to the server, it will play out the welcome prompt that “Welcome to continuous Kannada speech recognition system”. Followed by, the server will ask the user to speak the continuous Kannada speech sentence. Once the user speaks out the sentence, the spoken sentence will be recorded in the server and preprocessed (enhanced) using proposed noise elimination algorithm. Then, the server automatically checks the enhanced speech sentence text format in least WER ASR models. If the model exists for that particular enhanced speech sentence, then it will be recognized; otherwise, the server will ask the user to speak the speech sentence again (2nd time). If the speech sentence is recognized, then the server will ask the user that “Do you want to check the system with other continuous speech sentence?”. If the user says “yes”, then it will go to first stage and repeats the procedure, else it will be hanged up. If none of the speech sentences are recognized, then the server says “Sorry!! Try after sometime!!” at 3rd time is depicted in Fig. 7. The developed continuous Kannada SQS is tested by 500 speakers/users. Table 13 shows the online assessment of newly developed SQS by users under field conditions.

## 4 Conclusion

The design and implementation of a robust continuous Kannada ASR system by noise reduction technique have been described in this work. The various types of noises added in the collected continuous Kannada speech data have an adverse effect on

the entire ASR system performance in the earlier SQS. Therefore, we have developed a robust noise reduction technique for corrupted continuous Kannada speech data enhancement. The implementation procedural step of the proposed noise reduction algorithm was explained in detail. The proposed technique was applied to both training and testing speech dataset. For completeness of the current work, previously reported continuous Kannada SQS has been briefly depicted with its experimental results and analysis. In an end-to-end SQS, the proposed noise elimination algorithm was introduced before the speech feature extraction part. Once the test data are received from the speaker in SQS, the noise reduction algorithm has reduced various types of noises and recognized the continuous Kannada speech sentences. With the benefits of the proposed speech enhancement technique, Kannada language resources and Kaldi recipe, the obtained WER for enhanced continuous Kannada speech data was 2.23% by the combination of the proposed enhancement technique and TDNN. A better improvement is observed in minimizing the WER of 1.87% compared to earlier developed SQS for degraded continuous Kannada speech data. The online testing of developed current SQS (for enhanced speech data) is also done from 500 speakers, which revealed that there is an approximate match with online and offline accuracies of speech recognition.

## References

1. F. Albu, N. Dumitriu, L.D. Stanciu, Speech enhancement by spectral subtraction, in *Proceedings of International Symposium on Electronics and Telecommunications*, pp. 78–83 (1996)
2. M. Berouti, M. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 208–211 (1979)
3. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **27**(2), 113–120 (1979)
4. O. Cappé, Elimination of the musical noise phenomenon with the Ephraim, Malah noise suppressor. *IEEE Trans. Speech Audio Process.* **2**(2), 346–349 (1994)
5. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984)
6. N.W. Evans, J.S. Mason, W.M. Liu, B. Fauve, An assessment on the fundamental limitations of spectral subtraction, *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings.* **1**, 145–148 (2006)
7. H. Gustafsson, S.E. Nordholm, I. Claesson, Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Trans. Speech Audio Process.* **9**(8), 799–807 (2001)
8. H.G. Hirsch, D. Pearce, The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions, Automatic speech recognition: challenges for the new Millennium ISCA tutorial and research workshop (2000)
9. Y. Hu, P.C. Loizou, Evaluation of objective measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **16**(1), 229–238 (2008). <https://doi.org/10.1109/TASL.2007.911054>
10. Y. Hu, P.C. Loizou, Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun.* **49**, 588–601 (2007)
11. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **16**(1), 229–238 (2008)
12. ITU, Perceptual evaluation of speech quality (PESQ) and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation p. 862 (2001)
13. S. Kamath, P.C. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Process* (2002)

14. N. Kitaoka, S. Nakagawa, Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task. In: Seventh International Conference on Spoken Language Processing, ICSLP2002, pp. 477–480. Denver, Colorado, USA (2002)
15. B. Kumar, Comparative performance evaluation of greedy algorithms for speech enhancement system, *Fluctuation and Noise Letters*, World Scientific, vol. **20**(2) (2020)
16. S. Kumar, B. Kumar, N. Kumar, Speech enhancement techniques: a review. *Rungta Int. J. Electr. Electron. Eng.* vol. **1**(1), (2016)
17. P.S. Kumar, T.G. Yadava, H.S. Jayanna, Continuous Kannada speech recognition system under degraded condition. *Circuits Syst. Signal Process.* **39**(1), 391–419 (2019)
18. H. Liu, L. Zhao, A speaker verification method based on TDNN-LSTMP. *Circuits Syst. Signal Process.* **38**, 4840–4854 (2019)
19. P. Lockwood, J. Boudy, Experiments with a non-linear spectral subtractor (NSS) hidden Markov models and the projections for robust recognition in cars. *Speech Commun.* **11**, 215–228 (1992)
20. P.C. Loizou, Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. *IEEE Trans. Speech Audio Process.* **13**(5), 857–869 (2005)
21. P.C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007)
22. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **9**(5), 504–512 (2001)
23. A. Papoulis, S. Pillai, *Probability random variables and stochastic processes*, 4th edn. (McGraw-Hill Inc, New York, 2002)
24. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlce, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, The Kaldi Vesely, speech recognition toolkit, in *Proceedings IEEE, Workshop on Automatic Speech Recognition and Understanding* (US, Hilton Waikoloa Village, Big Island, Hawaii), p. 2011 (2011)
25. L.R. Rabiner, Applications of voice processing to telecommunications. *Proc. IEEE* **82**, 199–228 (1994)
26. S. Shahnawazuddin, K.T. Deepak, B.D. Sarma, A. Deka, S.R.M. Prasanna, S. Rohit, Low complexity on-line adaptation techniques in context of Assamese spoken query system. *J. Signal Process. Syst.* **81**(1), 83–97 (2015)
27. S. Shahnawazuddin, K.T. Deepak, D. Abhishek, I. Siddika, S.R.M. Prasanna, S. Rohit, Improvements in IITG Assamese spoken query system: background noise suppression and alternate acoustic modeling. *J. Signal Process. Syst.* **88**(1), 91–102 (2017)
28. N. Virag, Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Process.* **7**(3), 126–137 (1999)
29. M.R. Weiss, E. Aschkenasy, T.W. Parsons, Study and the development of the INTEL technique for improving speech intelligibility, in *Technical Report NSC-FR/4023*, (Nicolet Scientific Corporation, 1975)
30. T.G. Yadava, H.S. Jayanna, A spoken query system for the agricultural commodity prices and weather information access in Kannada language. *Int. J. Speech Technol.* **20**(3), 1–10 (2017)
31. T.G. Yadava, H.S. Jayanna, Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing. *Int. J. Speech Technol.* **22**(3), 639–648 (2018)
32. T.G. Yadava, H.S. Jayanna, Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling. *Int. J. Speech Technol.* **23**(1), 149–167 (2020)
33. T.G. Yadava, B.G. Nagaraja, H.S. Jayanna, Speech enhancement and encoding by combining SS-VAD and LPC. *Int. J. Speech Technol.* **24**, 165–172 (2021)
34. L. Yang, P.C. Loizou, A geometric approach to spectral subtraction. *Speech Commun.* **50**(6), 453–466 (2008)
35. N.B. Yoma, F.R. McInnes, M.A. Jack, Improving performance of spectral subtraction in speech recognition using a model for additive noise. *IEEE Trans. Speech Audio Process.* **6**(6), 579–582 (1998). <https://doi.org/10.1109/89.725325>