# Performance Enhancement of ASR System using Noise Reduction Technique as Pre-processing step

Anand H. Unnibhavi

Assistant professor, Department of Electronics and Communication Engineering
Basaveshwar Engineering College Bagalkot-587102, India
anandhu.rampur@gmail.com

D. S. Jangamashetti

Professor, Department of Electrical and Electronics Engineering
Basaveshwar Engineering College Bagalkot-587102, India
asdj1229@gmail.com

Shridhar S. Kuntoji

Professor, Department of Electronics and Communication Engineering
Basaveshwar Engineering College Bagalkot-587102, India
shridhar.ece@gmail.com

**Abstract**

**Most of the current Automatic Speech Recognition Systems give acceptable recognition accuracy for clean speech samples, but performance degrades when exposed to noisy situations. To overcome this, speech enhancement technique is used. In our earlier work Neutral Kannada Automatic Speech Recognition, word efficiency and sentence accuracy are 94.65% and 90% under clean speech environment. The same Recognition System when tested under Additive White Gaussian Noise and babble noise sources, the performance degrades. The relative word recognition accuracy degrades by 17.83% and sentence accuracy degrades by 21% under Additive White Gaussian Noise environment. The relative word recognition accuracy degrades by 17.56% and sentence accuracy by 21% in the presence of babble noise condition. In this work, to increase performance of Neutral Kannada Automatic Speech Recognition System noise reduction algorithm a preprocessing technique is proposed. It is based on conjugate symmetry of DFT. The improvement witnessed is about of 4% to 5%. Speech enhancement algorithm is evaluated using objective measures like Increment in segmental SNR, Log – likelihood ratio and Itakura – Saito distance.**

***Keywords*: ASR; DFT; AWGN; SNR; NKASR.**

## 1. Introduction

In the previous work Kannada ASR is implemented and tested for north Kannada accent. In this chapter the system will be exposed and evaluated different noise environment. Speaker recognition, videoconferencing, speech transmission via communication channels, speech-based biometric systems, mobile phones, hearing aids, voice conversion, and other applications all require speech augmentation. The development of voice augmentation algorithms

relies heavily on pattern mining technologies [1] In applications involving voice and speech, sound recording, cellular phones, hands-free communication, teleconferencing, hearing aids, and human-machine interfaces, a spoken signal of importance picked up via microphone sensor has always been jumbled with background noise. The statistical properties and spectrum of the desired voice signal can be severely contaminated by noise, depending on its amount. Noise not only introduces additional frequency components, but it also obscures a large percentage of the time varying spectral range of the intended speech. If the level of noise is not too high, one can still notice the relevant data buried in the intended speech signal but, if the intensity of noise is excessive, comprehension of the desired speech may become difficult. Many computer-based speech and speaker recognition, speech coding, and mobile communications systems are affected by noise [2]. Speech enhancement is a necessary condition in the realm of speech signal processing. The primary idea behind voice enhancement is reduction of background sound [3]. To filter and eliminate additive noise from a noisy voice signal, a variety of

algorithms based on classical techniques and machine learning have been used. The performance of ASR systems diminishes drastically in harsh settings, severely limiting the adoption of speech recognition applications. The degradation of ASR systems in noisy environments is caused by a mismatch among clean-trained acoustical models and noisy tested speech features provided to the voice recognition engine. Two general approaches of minimizing this mismatch are being offered. The first stage is to adapt the acoustical model to the noisy testing feature; the second step is to eliminate the noise from the noisy testing feature prior to recognition. Enhanced speech has been shown to improve human listener intelligibility and quality, and feeding it straight to ASR systems improves recognition accuracy [4, 5].

Existing speech enhancement algorithms are based on.

- Spectral subtraction: these techniques only work with the magnitude spectrum. It is depends on the noise spectral estimator, which is computed during the noisy speech signal's absence component.
- Subtraction of power spectral data these methods work by subtracting measured noise spectrum estimation from an audio signal power spectrum.
- Multiband spectral subtraction methods are used to remove colored noise.
- Phase aware speech enhancement: Noise induced phase deterioration is considered and accounted for in the phase spectrum.
- Speech improvement depending on evaluation: The Minimum Mean-Square Error Short-Time Spectral Amplitude estimator is used to enhance speech.
- Wiener filtering: This procedure is based on estimating the best filter from loud speech [6], and the filter is used for spectral subtraction.

Understanding and modeling individual diversity in spoken language is a crucial challenge for current research in ASR systems. Individuals have distinct speaking styles that are influenced by a variety of characteristics such as dialect and accent, as well as socioeconomic status, is all factors to consider. Individual differences can be difficult to model in large-scale speaker independent ASR systems that are designed to process input from any version of a language.

Different methods adopted to improve performance of ASR are:

- To boost the effectiveness of ASR, parallel phone recognition is accompanied by Language Modeling [7].
- Modeling of pronunciation dictionaries was utilized to reduce recognition problems caused by flaws in pronunciation. Each accent type requires a large corpus [8].
- Different automatic accent identification methods for geographically proximate dialects, as well as detecting the precise phonemes that are significant in a certain dialect recognition task and removing those that are not contributing in dialect recognition [9].
- An acoustic model that is accent independent and has been trained on data from all accents performs well on all data [10].
- Country-based automatic selection of the speech recognizer increases the accuracy [11, 12].

This research is carried out on evaluation of the performance of Neutral Kannada Automatic Speech Recognition (NKASR) systems on several noise scenarios and improving the performance of ASR using speech enhancement as preprocessing tool.

This work presents an evaluation of the performance of Neutral Kannada Automatic Speech Recognition (NKASR) systems on several noise scenarios and improving the performance of ASR using speech enhancement as preprocessing tool. The paper is organized as follows : Section 2 describes the proposed work, Section 3 deals with the details of performance evaluation of proposed noise reduction algorithm, Section 4 gives the experimentation details of the proposed method, Section 5 presents the results of ASR experiments and Section 6 discusses Conclusion.

## 2. Proposed work

In our earlier work [13] “triphone model based novel Kannada continuous speech recognition system using Kaldi”, was implemented for neutral Kannada (which is understood by all people in Karnataka irrespective of the region) ASR. The maximum recognition accuracy obtained for NKASR was 90%. The noise free NKASR performance is evaluated for the following conditions:

- noisy speech samples as input.
- noisy speech samples are input and with noise reduction technique as a preprocessing operation.

### *2.1 Proposed Enhancement Algorithm*

Speech recognition systems can achieve good recognition results in relatively quiet and intervention conditions. However, because the real-world scenario is so complex, achieving the ideal performance level of voice recognition is difficult [14]. To remove noise embedded in speech signal enhancement technique is employed and is based on the Complex Conjugate Symmetry algorithm of the Short-Time Fourier Spectrum (STFT), and its functioning is described as the magnitude of STFT of noisy speech is kept as it is while the phase is modified. Modified spectrum of speech is obtained by combining unchanged magnitude spectrum and modified phase spectrum. This modification results into cancellation of low energy components (noise) more than the high-energy (speech) components after speech reconstruction. The basic assumption of the proposed method is energy of speech components is more than that of noise components. The noisy speech signal x(n) is real; hence its DFT obeys conjugate symmetry [15]. The degree of cancellation or reinforcement of imaginary parts is controlled by modifying their phase in particular through their imaginary parts. This results in reduction of noise, which is essential in hearing aids as a front-end signal processing operation. Same is shown in Fig. 1.
Three phases of processing are included in the suggested technique.

- Analysis
- Spectrum modification
- Synthesis stage.

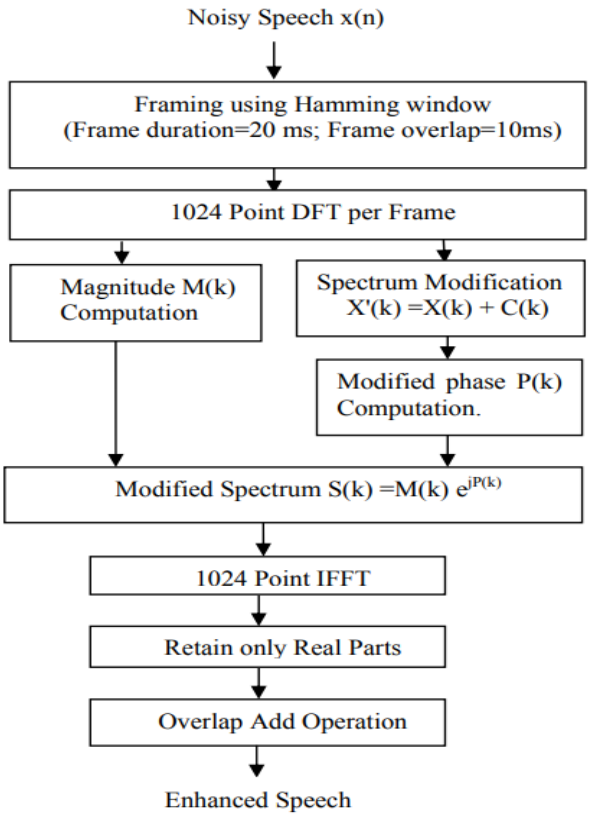


Fig. 1. Speech enhancement algorithm

Consider an additive noisy speech signal

$$x(n) = s(n) + y(n) \quad (1)$$

$x(n)$, $s(n)$, $y(n)$ are discrete noisy speech signal, clean speech signal and noise respectively. Since speech can be assumed to be quasi-stationary over 20–40 ms frame duration. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal is given by.

$$X(n,t) = \sum_{n=-\infty}^{\infty} x(m)w(n-m)e^{-\frac{j2tm}{N}} \quad (2)$$

Where $t$ denotes the $t^{th}$ discrete frequency of $N$ uniformly spaced frequencies and $w(n)$ is an analysis window function. Here Hamming window of 20 - 40 ms duration is employed.

The input noisy speech signal $x(n)$ is real, hence its DFT obeys conjugate symmetry i.e., $X(t) = X^*(N-t)$. IDFT of $X(t)$ results into original noisy speech signal $x(n)$ due to cancellation of imaginary parts of complex conjugate terms. But the degree of cancellation or reinforcement of complex conjugates can be controlled by modifying their phase [14]. A constant $\alpha(t)$, which is given by

$$\alpha(t) = c; \ 0 \le t \le \frac{N}{2} \quad (3)$$

$$\alpha(t) = -c; \ \frac{N}{2} \le t \le N-1 \quad (4)$$

$\alpha(t)$ is anti symmetric about the frequency $F_s/2$ and c is a real valued constant [15]. The noisy speech signal STFT $X(n, t)$ is modified as

$$L_a(t) = L(t) + a(t) \quad (5)$$

Computed $L_\alpha(t)$ modified phase and combined with magnitude of original noisy speech signal to get modified complex spectrum given by

$$L_m = L(t)e^{i\angle L_a(t)} \quad (6)$$

During the process of signal synthesis (IDFT) the conjugates sum together to result into a real signal due to cancellation of their imaginary parts. The degree of cancellation or summation of these complex conjugates can be controlled by modifying their phase. The above process can be visualized using signal vector analogy [16]. Considering a pair of complex conjugate numbers and having same magnitude,

$$M = \sqrt{L^2 + R^2} \quad (7)$$

and phase angles are,

$$\emptyset_1 = \tan^{-1}(R/L) \quad (8)$$

$$\emptyset_1^* = \tan^{-1}(-R/L) \quad (9)$$

These complex conjugate numbers are modified as,

$$C_{21} = L + j\,R + \text{ and } C_{21}^* = L - j\,R - C \quad (10)$$

The resulting phase angles are,

$$\emptyset = \tan^{-1}(R/L + C) \quad (11)$$

$$\emptyset^* = \tan^{-1}(-R/L - C) \quad (12)$$

Combining the magnitude M and modified phase as in equations (11), (12). The new complex conjugate numbers which can be expressed in polar form as

$$C_P = \sqrt{(L^2 + R^2)}\, e^{i\tan^{-1}(R/L+C)} \quad (13)$$

$$C_P{}^* = \sqrt{(L^2 + R^2)}\, e^{i\tan^{-1}(-R/L-C)} \quad (14)$$

The resultant of above two complex conjugate numbers is given by

$$C_{PR=} 2\sqrt{(L^2 + R^2)}\,; if\ C \ll M \quad (15)$$

The resultant obtained in above equation (15) is same as resultant of original complex numbers given in equation (7).

$$C_1 = L + jR = \sqrt{(L^2 + R^2)}\, e^{i\tan^{-1}(R/L)} \quad (16)$$

$$C_1{}^* = L - jR = \sqrt{(L^2 + R^2)}\, e^{i\tan^{-1}(-R/L)} \quad (17)$$

Resultant of above two complex numbers is given by

$$C = 2\sqrt{(L^2 + R^2)} \quad (18)$$

It is proved that from equation (15) and (18) $C_{PR} = C$. This implies that the phase modification due to $C$ has very negligible effect on the spectral components having magnitude more than magnitude of $C$. That is to say spectral components having magnitude more, nothing but speech components remain unaltered after phase modification as described above. Speech components remain same after phase modification. When C>>M the results are different and are explained below.

$$C_2 = U + jV \quad (19)$$

$$C_2^* = U - jV \quad (20)$$

Both having the same magnitude given by

$$M_2 = \sqrt{(U^2 + V^2)} \quad (21)$$

And phase angles

$$\emptyset_2 = \tan^{-1}(V/U) \quad (22)$$

$$\emptyset_2^* = \tan^{-1}(-V/U) \quad (23)$$

The complex conjugates phase angles are modified as

$$C_3 = U + jV + C \quad (24)$$

$$C_3^* = U - jV - C \quad (25)$$

$$\emptyset_3 = \tan^{-1}(V/U + C) \quad (26)$$

$$\emptyset_3^* = \tan^{-1}(-V/U - C) \quad (27)$$

Combining the magnitude of equation (21) and phase angles of equations (25), (26) results in new complex numbers and expressed in polar form as

$$C_2U = \sqrt{(U^2 + V^2)e^{i\tan^{-1}(V/U+C)}} \quad (28)$$

$$C_2U^* = \sqrt{(U^2 + V^2)e^{i\tan^{-1}(-V/U-C)}} \quad (29)$$

The resultant of above complex number is given by

$$M_{R2} = \sqrt{U^2 + V^2 + U^2 + V^2 + 2(U^2 + V^2)\cos(\theta)} \quad (30)$$

$$\theta = \tan^{-1}(V/U + C) + \tan^{-1}(-V/U - C) \quad (31)$$

If $C$ >>M2 then equation (30) can be written as

$$C_{R2} = \sqrt{2(U^2 + V^2)}\left[1 + \cos\left(\tan^{-1}\left[\frac{2VC}{C^2 - V^2}\right]\right)\right] \quad (32)$$

$$C_{R2} = \sqrt{2(U^2 + V^2)} \quad (33)$$

From equation (21) and (33) it is clear that the phase modification C has considerable effect on the spectral component for C << M2, i.e., the spectral component having magnitude less then C gets suppressed more after phase modification resulting in speech enhancement. The final outcome is, a particular value of C induces a definite value of phase modification in conjugate symmetrical spectral components leading to suppression of small spectral components (noise) and keeping the larger spectral components (speech) unaltered.

## 3. Performance Evaluation of proposed noise reduction algorithm

In this section, methods to evaluate the performance of the above speech enhancement technique procedure are discussed. The evaluation method used is Objective measures. The different objective measures are i) Log-likelihood ratio (LLR) ii) Itakura- saito distance (ISD) and iii) Increment in Segmental SNR [17, 18].

- **Log Likelihood Ratio**

  LLR is calculated by

$$d_{LLR}(a_e a_c) = log\left(\frac{a_e R_c a_e^T}{a_e R_c a_e^T}\right) \quad (34)$$

  Where,
  $a_c$ is linear predictive coding (LPC) vector of clean speech signal
  $a_e$ is LPC vector of enhanced or processed speech signal
  $R_c$ is autocorrelation matrix of clean speech signal

- **Itakura – Saito distance (ISD)**

  ISD compares the position of peaks and valleys of enhanced speech with respect to enhanced speech and is calculated by

$$D_{ISD}(a_e a_c) = \frac{G_C}{G_e}\left(\frac{a_e R_c a_e^T}{a_e R_c a_e^T}\right) + \log\left(\frac{G_e}{G_c}\right) - 1 \quad (35)$$

  Where,
  $G_e$ is LPC gain of processed speech
  $G_c$ is LPC gain of clean speech respectively.

- **Increment in Segmental SNR**

It is the difference between signal to noise ratio of enhanced speech and noisy speech SNR. It is calculated for every frame and average of the entire frame is taken and is calculated by,

$$SNR_{seg} = \left(\frac{10}{l}\right)\sum_{i=0}^{l-1} log_{10}\left(\frac{\sum_{n=Ni}^{Ni+N-1} x^2(n)}{\sum_{n=Ni}^{Ni+N-1}\left(x(n) - x(\hat{n})\right)^2}\right) \qquad (36)$$

Where,
$N$ is Frame length
$l$ is Number of frames of $x(n)$
$x(n)$ is noisy speech signal
$x(n)$ is processed speech signal.

## 4. Experimentation

In this research work the performance of NKASR is evaluated when exposed to AWGN and Babble noise. The NKASR performance degrades under noisy condition. To improve the performance of NKASR the noisy speech samples are enhanced with speech enhancement technique which increases the accuracy even when tested under noisy condition.

### *4.1 Performance of ASR under noisy speech samples*

The extraction of acoustic features that are most relevant about speech is a predominant process at the front-end of all speech technologies (recognition systems, coding schemes, hearing prosthesis, and so on). However, because continuous speech has a high level of redundancy, this procedure provides a significant barrier to all of these systems. In our previous work [13] NKASR is implemented using Kaldi software. The wave surfer tool is used to capture a total of 1200 sentences (4 speakers x 300 sentences=1200 sentences). The spoken sentences are sampled at a rate of 16 kHz with a depth of 16 bits. A total of 1100 audio files are used for training HMM (Hidden Markov Model) and 100 audio files are used for testing. Recognition accuracy obtained for neutral Kannada is 90%.

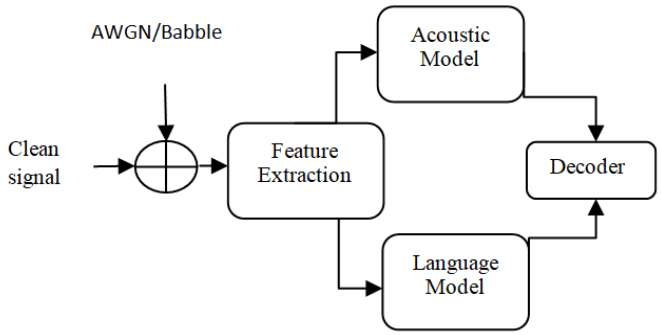


Fig. 2 NKASR system with noisy input

Two types of noise signal considered are: AWGN and Babble.
NKASR as shown in Fig. 2 is tested with 88 clean audio files and 12 clean audio files which are added with Additive White Gaussian Noise (AWGN) of 0dB, 5dB, 10dB and 15dB SNR. The performnce of NKASR gets degraded as the NKASR is trained with noise free audio files and tested for noisy speech files, i.e. mismatch between training and testing conditions. Theoretically the addition of AWGN noise, norm (length) of cepstrum coefficients reduces as SNR decreases [17]. To maintain robustness of NKASR, MFCC feature extracting method is used to extract features and is more resistance to noisy environment than conventional LPC cepstrum coefficients. Another feature normalization technique used is, cepstral mean and variance normalization (CMVN) which retains the robustness of NKASR

Similarly NKASR is tested with 88 clean audio files, and 12 clean audio files are clubbed with babble noise (Restaurant) of 0dB, 5dB, 10dB and 15dB SNR. This noise is not uniformly distributed in clean signal, instead babble noise settles at all informative areas (as clean signal is audio signal and babble signal is also audio signal hence babble formants overlaps with clean signal formants) of clean signal so performance of NKASR gets degraded to large extent in comparison with AWGN noise.

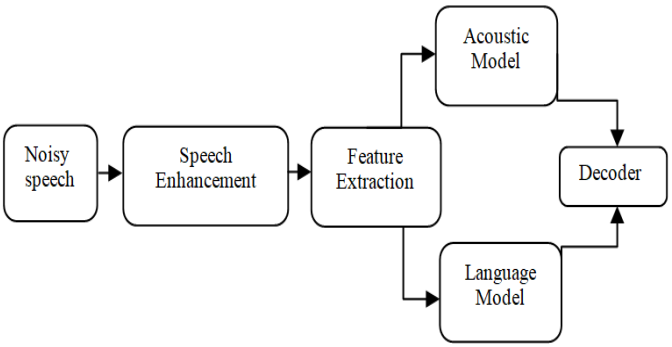


Fig. 3 NKASR system with Speech Enhancement

### *4.2 Noisy speech input with noise reduction technique as a preprocessing operation*

In section 4.1, when noise (AWGN, Babble noise) is added to clean signal, the performance of NKASR degrades due to the mismatch between training and testing conditions. To increase the performance of NKASR, speech enhancement technique as a preprocessing operation is adopted before speech recognition as shown in Fig. 3. Speech enhancement techniques minimize the noise and enhance the parameter vectors of the clean speech embedded in noise. Speech enhancement technique used is based on "complex conjugate symmetry presence of DFT". In this algorithm a particular value of C modifies the phase of speech signal in conjugate symmetrical spectral components that suppress small spectral components (noise), without affecting larger spectral components (speech).

| | | SNR Values | | | |
|---|---|---|---|---|---|
| SL. NO | Type of Noise (db), | 0 | 5 | 10 | 15 |
| 1. | Babble | 9.50 | 3.8 | 1.90 | 0.85 |
| 2. | Train | 6.00 | 2.9 | 1.45 | 0.75 |
| 3. | Airport | 3.50 | 1.25 | 1.00 | 0.75 |
| 4. | Exhibition | 4.00 | 3.70 | 2.30 | 1.60 |
| 5. | Restaurant | 3.00 | 2.50 | 1.80 | 1.10 |
| 6. | Street | 5.75 | 4.00 | 2.20 | 1.10 |
| 7. | Station | 6.00 | 4.00 | 2.00 | 0.85 |
| 8. | AWGN | 3.50 | 2.00 | 1.05 | 0.55 |

The experimental values of C are shown in Table.1.

### *4.3 Algorithm Procedure*

The clean speech samples after adding with AWGN and Babble noise (Restaurant) with SNR of 0, 5, 10 and 15dB are sampled at 16 kHz, and framed with speech frame duration of 20ms with 10ms overlap. Hamming window multiplies each frame. The FFT is computed with 256 points. The first 128-point FFT is altered by adding a constant C, and the other 128-point FFTs are altered by subtracting the same constant C. The audio SNR is represented by C. Phase of each modified DFT samples is calculated. Original DFT magnitudes and adjusted phase values are combined to create the new DFT samples. In the final step IFFT is applied and discrete time signals are found by overlap and add method. Finally, the speech enhancement algorithm is evaluated for objective tests like, Log likely hood ratio, Itakura saito distance and Increment in segmental SNR.

### *4.4 Result and analysis*

In our earlier work [13], word accuracy obtained is 94.65% and sentence accuracy is 90% under clean speech samples, In the work presented in this section NKASR performance is tested by noisy samples as input, 100 speech samples are used for validation, out of which 88 speech samples are clean and 12 speech samples are corrupted by i). AWGN of 0dB, 5dB, 10dB and 15dB noise, the same 88 clean and 12 corrupted speech samples are given as input to NKASR. Table 2 shows performance of NKASR in terms of word and sentence accuracy for clean and noisy speech signals. Word accuracy obtained are 94.65%, 76.82%, 77.09% and sentence accuracy obtained are 90%, 69%, 69% respectively for clean speech, AWGN noisy speech and babble noisy speech respectively. The performance of NKASR under AWGN noise with word accuracy is 76.82 %, sentence accuracy is 69%. The result shows degradation in performance of NKASR. i.e., the word accuracy degrades by 17.83% and sentence accuracy degrades by 21%.

Similarly, NKASR performance is tested for Babble (Restaurant) noise. Word accuracy obtained is 77.09 and sentence accuracy is 69% as shown in Fig.4. The word accuracy drops by 17.56%, and sentence accuracy drops by 21%. The performance of NKASR degrades is due to mismatch in training and testing environment (clean and noisy signal).

| Input to NKASR | Word accuracy | Sentence accuracy |
|---|---|---|
| Clean samples | 94.65 | 90 |
| AWGN + Clean samples | 76.82 | 69 |
| Babble + Clean samples | 77.09 | 69 |

Table 2. NKASR performance under clean samples, clean samples with AWGN and Babble noise

| SNR (dB) | Word accuracy for noisy speech (%) | Word accuracy for enhanced speech (%) |
|---|---|---|
| 0 | 75.70 | 81.01 |
| 5 | 76.82 | 81.01 |
| 10 | 76.82 | 79.65 |
| 15 | 75.14 | 81.56 |

Table 3. NKASR performance of word accuracy for Noisy speech (AWGN) and enhanced speech

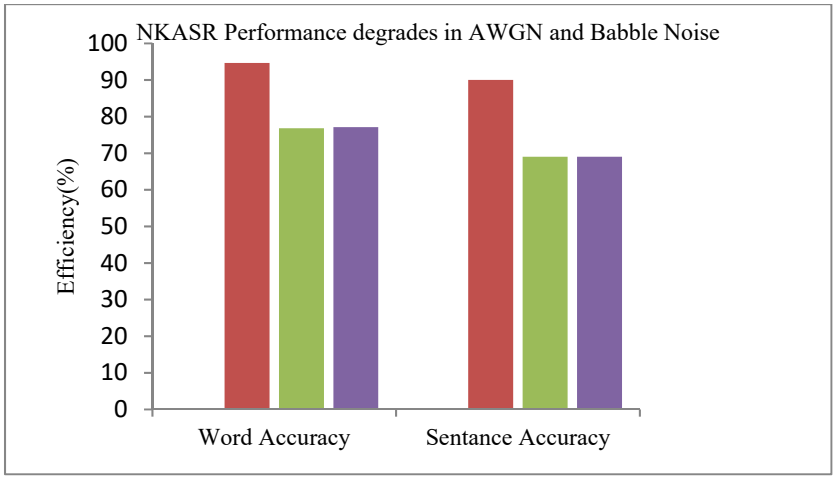


Fig.4. Performance degradation of NKASR system.

To improve the performance of NKASR, it is necessary to recover clean speech embedded in noise due to the addition of AWGN and Babble noise in the speech samples. A speech enhancement technique called conjugate symmetry of DFT is used. The corrupted speech samples are passed through noise reduction algorithm. The resulting speech samples are called as enhanced speech and are given as input to NKASR.

Table 3 gives NKASR performance for word accuracy in case of noisy speech (AWGN) and enhanced speech signal. The word accuracies for noisy speech are 75.70%, 76.82%, 76.82%, 75.14% and for enhanced speech word accuracies are 81.01%, 81.01%, 79.65%, 81.56% respectively for different AWGN noise level 0dB, 5dB, 10dB and 15dB. In case of AWGN noisy enhanced speech samples NKASR performance increases word accuracy to 81.56 % as shown in Fig. 5. WA and SA are functions of phoneme levels. For monophone modeling WER is comparatively more with respect to triphone modeling which is evident in the graphs shown.

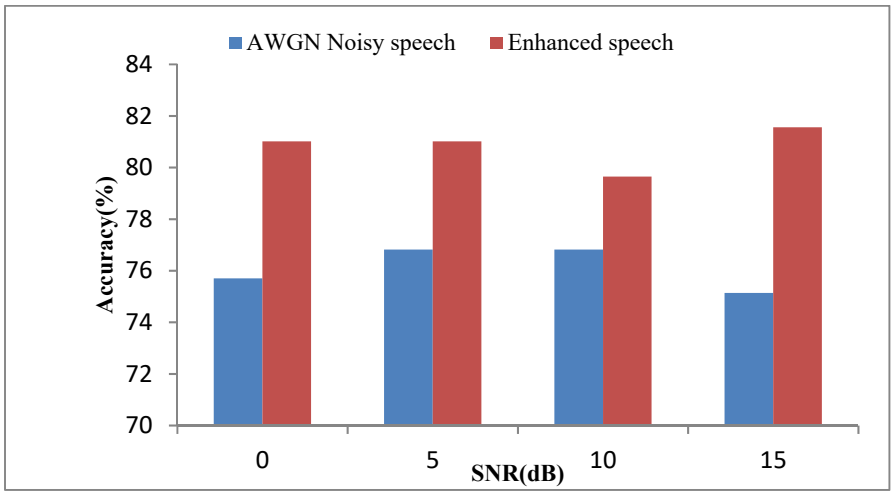


Fig.5. Word accuracy of NKASR for Noisy speech (AWGN) and enhanced speech

Table 4 gives NKASR performance for sentence accuracy in case of noisy speech (AWGN) and enhanced speech signal. Sentence accuracies for noisy speech are 68%, 69%, 69%, 67% and sentence accuracies for enhanced speech are 74%, 71%, 70%, 74% respectively for different AWGN noise level 0dB, 5dB, 10dB and 15dB. In case of AWGN noisy enhanced speech samples NKASR performance increases sentence accuracy to 74 % as shown in Fig. 6

| SNR (dB) | Sentence accuracy for noisy speech (%) | Sentence accuracy for enhanced noisy speech (%) |
|---|---|---|
| 0 | 68 | 74 |
| 5 | 69 | 71 |
| 10 | 69 | 70 |
| 15 | 67 | 74 |

Table 4. NKASR performance of sentence accuracy for AWG Noise and enhanced speech

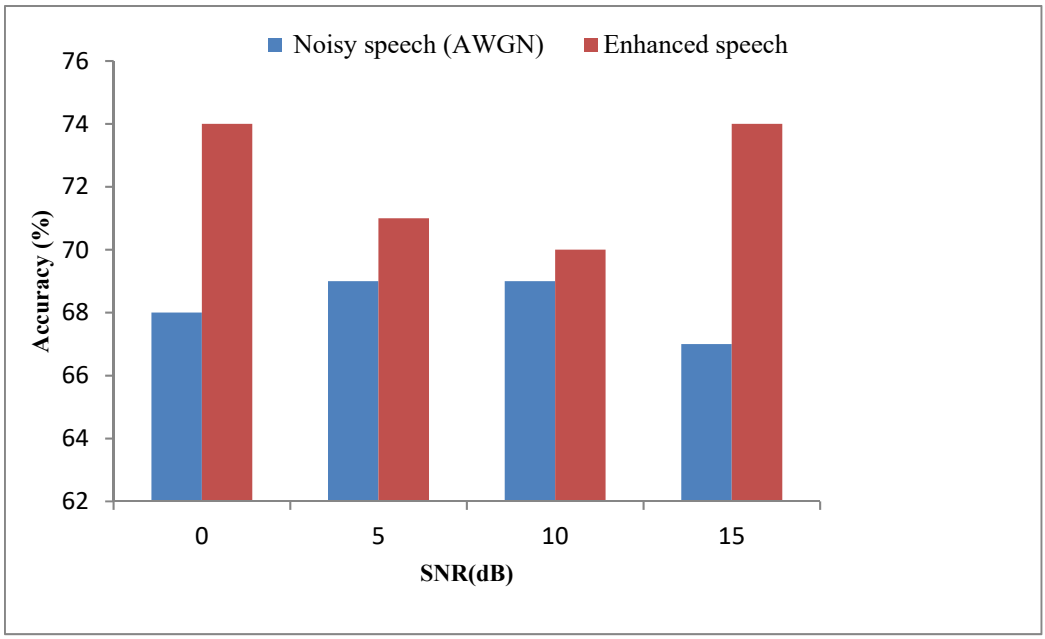


Fig.6. Sentence accuracy of NKASR for AWG Noise and enhanced speech

Table 5 gives NKASR performance for word accuracy in case of noisy speech (Babble) and enhanced speech signal. Word accuracies for noisy speech are 75.64%, 75.70%, 76.54%, 77.09% and word accuracies for enhanced speech are 80.45%, 80.45%, 77.65%, 77.65% respectively for different Babble noise level of 0dB, 5dB, 10dB and 15dB. In case of Babble noisy enhanced speech samples NKASR performance increases word accuracy to 81.84 % as shown in Fig. 7.

| SNR (dB) | Word accuracy for noisy speech (%) | Word accuracy for enhanced noisy speech (%) |
|---|---|---|
| 0 | 75.64 | 80.45 |
| 5 | 75.70 | 80.45 |
| 10 | 76.54 | 81.84 |
| 15 | 77.09 | 77.65 |

Table 5. NKASR performance for word accuracy under Babble noise and enhanced speech

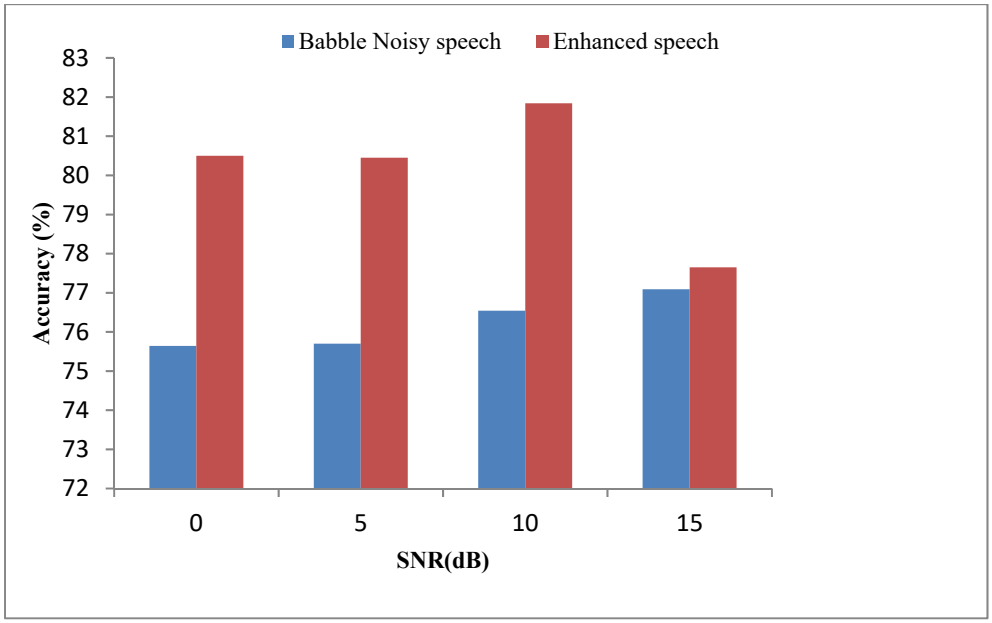


Fig. 7. Word accuracy of NKASR for Babble noise and enhanced speech.

Table 6 gives NKASR performance for sentence accuracy in case of noisy speech (Babble) and enhanced speech signal. Sentence accuracies for noisy speech are 67%, 69%, 69%, 69% and sentence accuracies for enhanced noisy speech are 68%, 73%, 73%, 69% respectively for different Babble noise level 0dB, 5dB, 10dB and 15dB. In case of Babble noisy enhanced speech samples NKASR performance increases sentence accuracy to 73 % as shown in Fig. 8.

| SNR (dB) | Sentence accuracy for Noisy speech (%) | Sentence accuracy for Enhanced noisy speech (%) |
|---|---|---|
| 0 | 67 | 68 |
| 69 | 69 | 73 |
| 10 | 69 | 73 |
| 15 | 69 | 69 |

Table 6. NKASR performance for sentence accuracy under Babble noise and enhanced speech

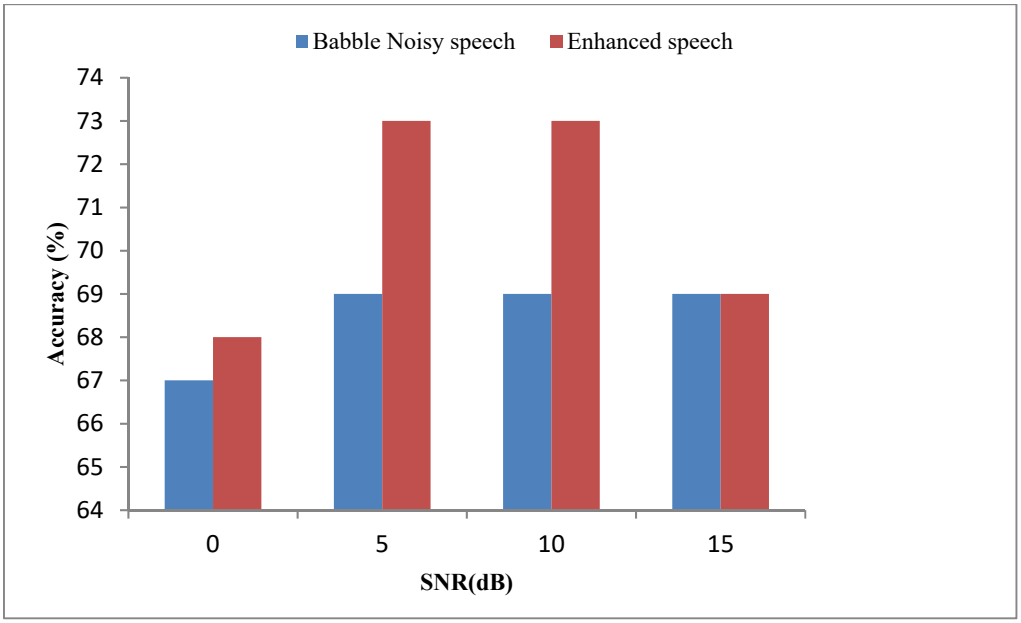


Fig. 8. Sentence accuracy of NKASR for Babble noise and enhanced speech

Table 7 gives the details of the relative performance improvement of NKASR system. The relative performance improvement is 6.17% of word accuracy, 7.24% of sentence accuracy after enhancement under AWG Noisy environment. Similarly 6.16% and 5.79% relative improvement under Babble noise condition.

| WA (%) | Enhanced WA (%) | (%) word relative improvement | SA (%) | Enhanced SA (%) | (%) sentence relative improvement |
|---|---|---|---|---|---|
| AWG Noise | | | | | |
| 76.82 | 81.56 | 6.17 | 69 | 74 | 7.24 |
| Babble Noise | | | | | |
| 77.09 | 81.84 | 6.16 | 69 | 73 | 5.79 |

Table 7 Relative performance improvement of NKASR System

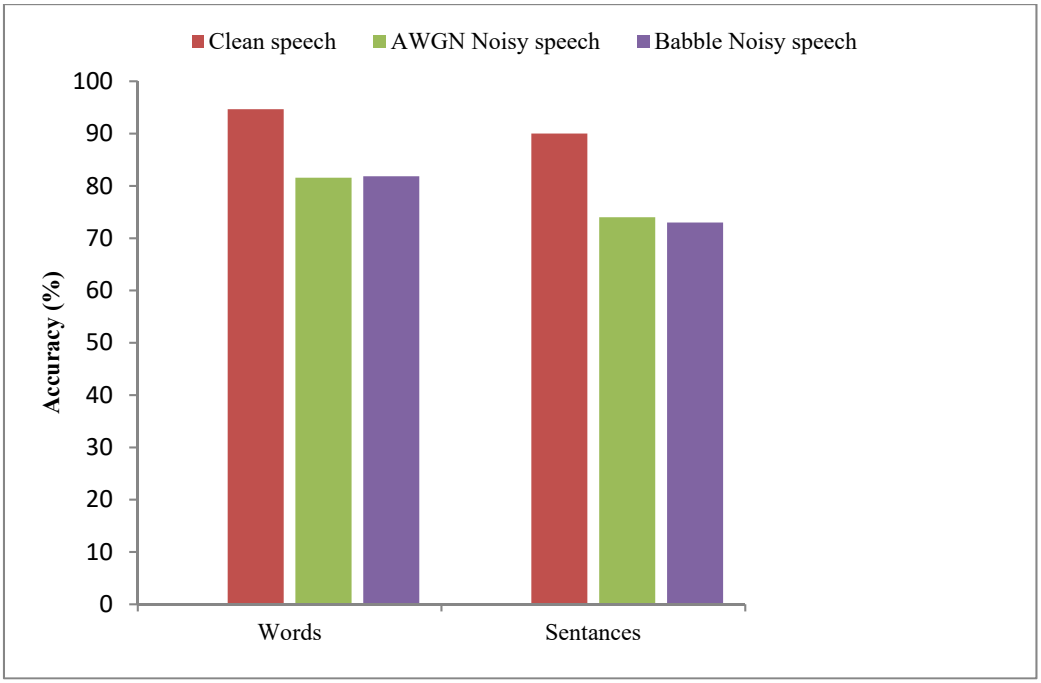


Fig. 9. Performance improvement of NKASR system.

From the above result, it is clear that the performance of NKASR improves with noise reduction algorithm as a preprocessing step. For AWGN and Babble enhanced speech word accuracy increases to 81.56 %, 81.84 % and sentence accuracy to 74 %, 73 % respectively as shown in Fig. 9.

### *4.4.1 Evaluation of Noise reduction Algorithm*

**1) Log-Likely hood ratio (LLR)**

- From the Fig. 10, it is observed that distortion induced is more in clean speech for 0db and 5db by AWGN so LLR value is more than 2, but for 10db, 15db LLR value is less than 2 distortion induced is less.
- From the Fig. 11, it is observed that distortion induced is less in clean speech for 0db, 5dB, 10dB, 15dB, by Babble noise so LLR value is less than 2.

Therefore, noise reduction algorithm performance is good and resulting in best speech recognition accuracy in presence of both AWGN and Babble noise.

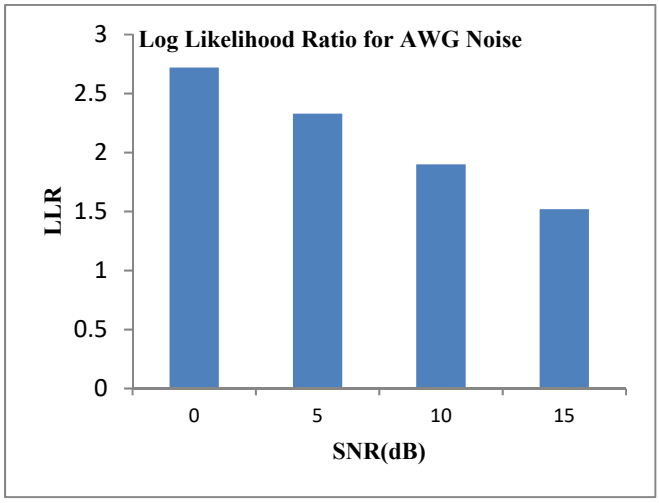


Fig. 10. Log likelihood ratio v/s SNR

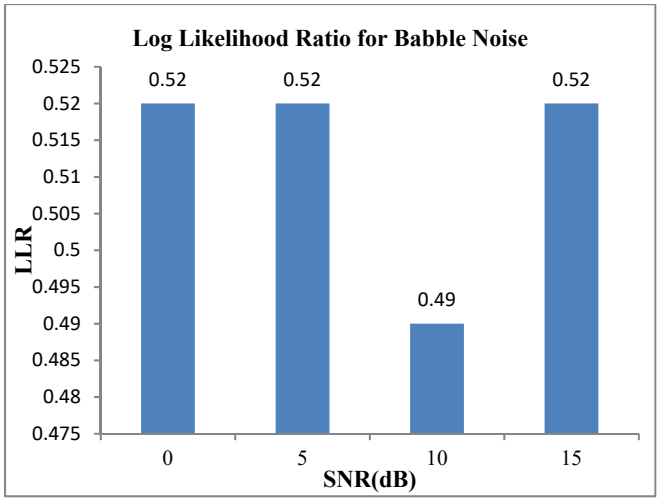


Fig. 11. Log likelihood ratio v/s SNR

### 2) Itakura – Saito Spectral Distance

- From Fig. 12, it is observed that ISD values obtained are less than 100 for AWGN noise and also for Babble noise.
- Maximum ISD value in Fig. 12 is 71.91 which is more when compared to Fig. 13 with ISD value of 21.4. More distortion is observed with AWGN noise than Babble noise. Therefore, the performance of the proposed noise reduction algorithm resulting in improved performance of NKASR.

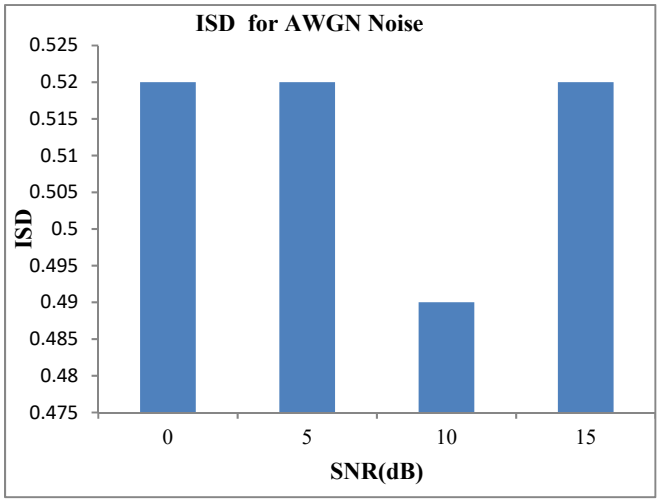


Fig. 12. Bar graphs of ISD for AWGN Noise

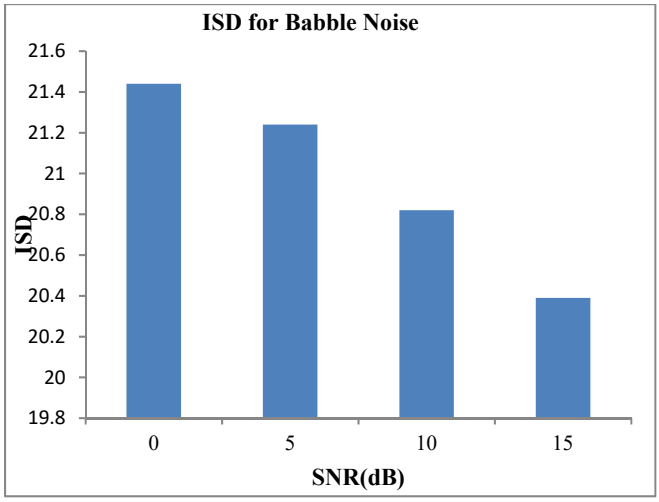


Fig. 13. Bar graphs of ISD for Babble Noise

### 3) Incremental Segmental SNR

- From Fig.14, it is observed that as SNR is increased under AWGN noisy condition there is growth in Incremental segmental SNR. Similarly, growth of Incremental segmental SNR is observed in Fig.15 under Babble noise condition.
- It is also observed that as SNR of the noisy speech signal increases, increment in segmental SNR decreases, because more signal strength and less noise, so the proposed noise reduction algorithm subtracts unnecessarily little noise along with useful data, this reduces the performance of the proposed noise reduction algorithm resulting in average growth of performance in NKASR. Therefore, overall, the performance of the suggested noise reduction method has improved, resulting in better NKASR performance.

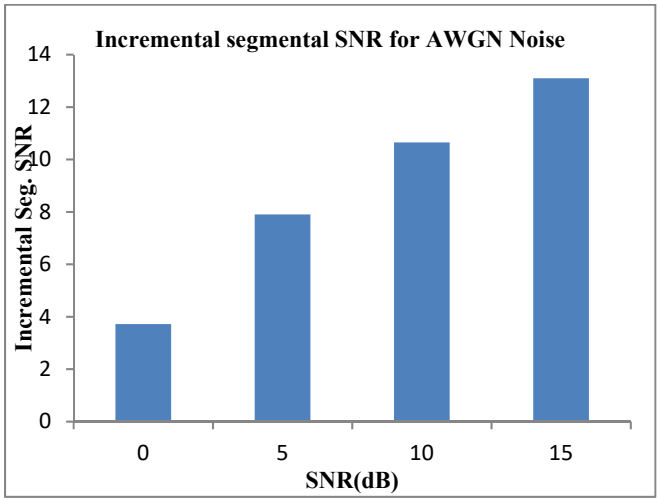


Fig. 14. Bar graphs of ISD for Incremental segmental SNR under AWGN Noise

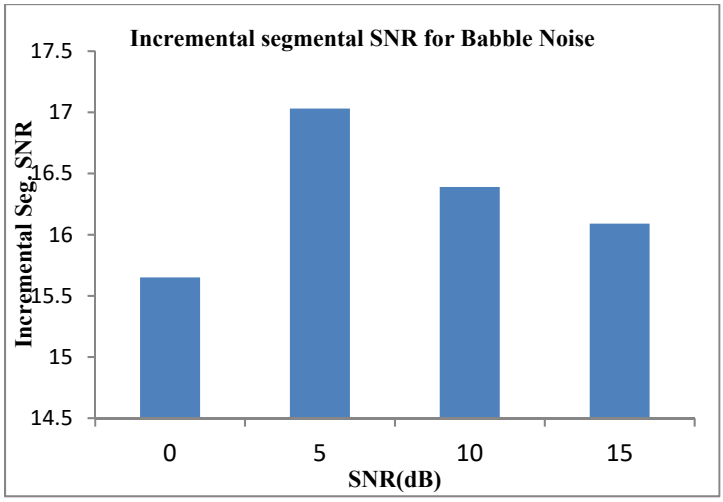


Fig. 15. Bar graphs of ISD for Incremental segmental SNR under Babble Noise

## 5. Conclusion

In our earlier work Neutral Kannada Automatic speech Recognition System (NKASR) was built. Word accuracy obtained is 94.65% and sentence accuracy is 90% under clean speech [13]. In this work when NKASR tested under AWGN, Babble noise the performance degrades. The performance of NKASR under AWGN noise with word accuracy drops down to76.82 %, sentence accuracy to 69 %. The relative word accuracy degradation is 17.83% and sentence accuracy degradation is 21%. Similarly NKASR performance when tested for Babble (Restaurant) noise word accuracy falls down to 77.09 %, sentence accuracy to 69%. The relative word accuracy degradation is 17.56%, and sentence accuracy degradation is 21%. To make NKASR robust to noisy condition a noise reduction algorithm "conjugate symmetry of DFT" is designed to improve the performance of NKASR, This algorithm recover clean speech embedded in noise due to the addition of AWGN and Babble noise in the clean speech samples. This algorithm reduces the mismatch between training and validation conditions. The proposed algorithm increases NKASR performance to 81.56 %, sentence accuracy to 74% under AWGN. Similarly in Babble noisy enhanced speech samples NKASR performance increased by 81.84% with word accuracy and sentence accuracy by 73%. The relative accuracy improvement between AWGN noisy and enhanced samples are 4.74% with word accuracy and 5% sentence accuracy. The relative improvement between Babble noise and enhanced samples are 4.75% with word accuracy and 4% sentence accuracy. From the above result it is clear that performance of NKASR improves with noise reduction algorithm as a preprocessing step. Also Noise reduction algorithm designed is made to undergo evaluation for objective measures like Increment in segmental SNR, Log – likelihood ratio and Itakura – saito spectral distance. In next work NKASR performance of the system with respect north Kannada accent will be improved.

## Conflict of interest

All authors declare no conflicts of interest in this paper.

## References

[1] Nabanita Das, Sayan Chakraborty, Jyotismita Chaki, Neelamadhab Padhy, Nilanjan Dey (Dec 2021), "Fundamentals, present and future perspectives of speech enhancement," International Journal Speech Technology, Vol. 24, No. 4, pp. 883–901, doi: 10.1007/s10772-020-09674-2.

[2] Khaled Daqrouq, Ibrahim N. Abu-Isbeih and Omar Daoud, Emad Khalaf (June 2010), 'An investigation of speech enhancement using wavelet filtering method', International Journal Speech Technology, Vol. 13, no. 2, pp. 101-115, doi: 10.1007/s10772-010-9073-1.

[3] Jacob Benesty, Jingdong Chen and Emanuel A. P. Habets (2012), 'Speech Enhancement in the STFT Domain', Springer Briefs in Electrical and Computer Engineering.

[4] Xiao Xiong and Xie lei (2006), 'Speech Enhancement with Applications in Speech Recognition', Semantic Scholar.

[5] Peidong Wang and Ke Tan (2019), 'Bridging the Gap Between Monaural Speech Enhancement and Recognition with Distortion-Independent Acoustic Modeling', Transactions on Audio, Speech, and Language Processing.

[6] Dash T. K., and Solanki S. S. (2017), 'Comparative Study of Speech Enhancement Algorithms and Their Effect on Speech Intelligibility'. Proceedings of the 2nd International Conference on Communication and Electronics Systems.

[7] Fadi Biadsy, 'Automatic Dialect and Accent Recognition and its Application to Speech Recognition', Doctor of Philosophy thesis in the Graduate School of Arts and Sciences Columbia University, 2011.

[8] Chao huang, Tao chen and Eric chang (2004), 'Accent Issues in Large Vocabulary Continuous Speech Recognition', International Journal of Speech Technology, pp. 141–153.

[9] Georgina Brown (June 2016), 'Automatic Accent Recognition Systems and the Effects of Data on Performance', Odyssey 2016, doi: 10.21437/Odyssey.2016-14.
[10] Dimitra Vergyri, Lori Lamel, Jean-Luc Gauvain (2010), 'Automatic Speech Recognition of Multiple Accented English Data', INTERSPEECH 2010, doi: 10.21437/Interspeech.2010-477.
[11] Mohamed G. Elfekya, Pedro Morenoa and Victor Sotob (2015), 'Multi-Dialectical Languages Effect on Speech Recognition', International Conference on Natural Language and Speech Processing.
[12] Hao Wu1 and Arun Kumar Sangaiah (2021), 'Oral English Speech Recognition Based on Enhanced Temporal Convolutional', Intelligent Automation & Soft Computing, Vol. 28, doi:10.32604/iasc.2021.016457
[13] Anand H.Unnibhavi, D. S. Jangamshetti and Shridhar K (2020), 'Triphone Model Based Novel Kannada Continuous Speech Recognition System using Kaldi Tool', International Journal of Innovative Technology and Exploring Engineering, Vol. 9, No.9.
[14] Kamil Wojcicki, Mitar Milacic, Anthony Stark, James Lyons, and Kuldip Paliwal (2008), 'Exploiting conjugate symmetry of the short – time Fourier spectrum for speech enhancement', IEEE Signal processing letters, Vol. 15, pp. 461- 464.
[15] Shaila Apte and Shridhar (2010), 'Speech Enhancement in Hearing Aids Using Conjugate Symmetry of DFT and SNR-Perception Models', International Journal of Computer Applications, Vol. 1, Issue 21, February 2010. doi: 10.5120/58-650.
[16] Naik D C, A Sreenivasa Murthy, Ramesh Nuthakki, 'A Literature Survey on Single Channel Speech Enhancement Techniques', International Journal of Scientific & Technology Research (IJSTR), Vol. 9 - Issue 3, March 2020 Edition.
[17] Yifan Gong (1995), "Speech Recognition in Noisy Environments: A Survey," Speech Communication, Vol. 16, No. 3, pp. 261-291.

## Authors Profile

Dr.Anand H. Unnibhavi completed his B.E in Electronics and communication Engineering from Vidya Vardhaka college of Engineering Mysore, affiliated to VTU Belagavi Karnataka and obtained his M.Tech in the area of Digital Electronics and Communcation system from Malnad College of Engineering Hassan , affiliated to VTU Belagavi Karnataka. Completed Ph.D in the area of Speech Processing. Areas of interest are Speech processing, Wireless network. Presently working as Assistant professor in the department of Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot,Karnataka, India.

Dr. Dakshayani S. Jangamshetti was born in Ilkal, Karnataka, India on 12th February 1964. She obtained her B.E (Electrical) degree from Karnataka University Dharwad in 1985 and M.Tech.(Instrumentation) Degree from IIT Kharagpur in 1989 & Ph.D (Speech Processing) from IIT, Mumbai in 2003. Her areas of interest include speech signal processing, Image processing, Microcontroller and signal systems. She won the "Outstanding IEEE Branch Counselor" award for the year 2014. Presently, she is a principal of Basaveshwar Engineering College (Autonomous), Bagalkot.

Dr.ShridharS.Kuntoji received his B.E (Electronics and Communication Engineering) degree from Gulbarga University Gulbarga, M.Tech (Digital Electronics and advanced communication) from NITK Surathkal, and Ph.D in the area of speech signal processing from Shivaji University Kolhapur in the year 1988, 2000 and 2014 respectively. He joined as Lecture in Electronics and Communication Engineering, Basaveshwara Engineering College Bagalkot, India in the year 1993, where he is currently working as Professor since 2014. He is currently involved in the research area of speech enhancement focusing on people suffering from hearing loss.