Thimmaraja G. Yadava* and H.S. Jayanna

Improvements in Spoken Query System to Access the Agricultural Commodity Prices and Weather Information in Kannada Language/Dialects

https://doi.org/10.1515/jisys-2018-0120
Received March 1, 2018; previously published online June 20, 2018.

Abstract: In this paper, the improvements in the recently developed end to end spoken query system to access the agricultural commodity prices and weather information in Kannada language/dialects is demonstrated. The spoken query system consists of interactive voice response system (IVRS) call flow, automatic speech recognition (ASR) models and agricultural commodity prices, and weather information databases. The task specific speech data used in the earlier spoken query system had a high level of background and other types of noises as it is collected from the farmers of Karnataka state (a state in India that speaks the Kannada language) under uncontrolled environment. The different types of noises present in collected speech data had an adverse effect on the on-line and off-line recognition performances. To improve the recognition accuracy in spoken query system, a noise elimination algorithm is proposed in this work, which is a combination of spectral subtraction with voice activity detection (SS-VAD) and minimum mean square error spectrum power estimator based on zero crossing (MMSE-SPZC). The noise elimination algorithm is added in the system before the feature extraction part. In addition to this, alternate acoustic models are developed using subspace Gaussian mixture models (SGMM) and deep neural network (DNN). The experimental results show that these modeling techniques are more powerful than the conventional Gaussian mixture model (GMM) – hidden Markov model (HMM), which was used as a modeling technique for the development of ASR models to design earlier spoken query systems. The fusion of noise elimination technique and SGMM/DNN-based modeling gives a better relative improvement of 7% accuracy compared to the earlier GMM-HMM-based ASR system. The least word error rate (WER) acoustic models could be used in spoken query system. The on-line speech recognition accuracy testing of developed spoken query system (with the help of Karnataka farmers) is also presented in this work.

Keywords: Noise elimination, IVRS, ASR, accuracy, spoken query system.

1 Introduction

Automatic speech recognition (ASR) system using an interactive voice response system (IVRS) is one of the important applications of speech processing [27]. The amalgamation of IVRS and ASR systems are called spoken query systems, which are used to decode the user input [10, 33], and the needed information is spread by the system. The recent advancement in the speech recognition domain is that the touch tones used in the earlier ASR systems have been completely removed. A spoken query system has been developed recently to access the agricultural commodity prices and weather information in Kannada language/dialects [32]. This work is an ongoing sponsored project by the Department of Electronics and Information Technology (DeitY), Government of India, targeted to develop user-friendly spoken query system to access the commodity prices and weather information by addressing the needs of Karnataka farmers. The developed spoken query system gives an option to the user/farmer to make his own query about any agricultural commodity prices and weather information over mobile/landline telephone network. The query, which is uttered by the user/farmer, is recorded, the price/weather information in the database is checked through ASR models, and the on-time

***Corresponding author: Thimmaraja G. Yadava,** Research Scholar, Panini Research Center, 3rd Floor, Department of ECE, Siddaganga Institute of Technology, Tumkur, Karnataka 572103, India, e-mail: thimrajyadav@gmail.com
H.S. Jayanna: Department of ISE, Siddaganga Institute of Technology, Tumkur, Karnataka, India

price/weather information of a particular commodity in a desired district is communicated through pre-recorded messages (voice prompts). The earlier spoken query system [32] is developed using the Gaussian mixture model (GMM)-based hidden Markov model (HMM). In this work, we demonstrate the improvements to the earlier spoken query system. A noise elimination algorithm is proposed, which is used to reduce the noise in the speech data collected under an uncontrolled environment. We have also investigated the two different acoustic modeling approaches reported latterly subjected to the spoken query system. The training and testing speech data used in Ref. [32] for the development of ASR models was collected from the farmers across the different dialect regions of Karnataka (a state in India speaks Kannada language) under real-time environment. Therefore, the collected speech data was adulterated by different types of background noises, which includes babble noise, background noise, car noise, vocal noise, and horn noise, etc., while the user/farmer making a query to the system also happens to have a high level of background noise. This totally decreases the entire spoken query system performance. To overcome this problem, we have introduced the noise elimination algorithm before the feature extraction part. This algorithm eliminates the different types of noises in both training and testing speech data. The removal of background noises leads to a good modeling of phonetic contexts. Therefore, an improvement in the on-line and off-line speech recognition accuracy is achieved compared to the earlier spoken query system.

The noise reduction in degraded speech data is a challenging task [20, 28]. The spectral subtraction (SS) method is commonly used for speech enhancement and is mainly associated with voice activity detection (VAD). To find the active regions of degraded speech signal, VAD is used [29]. The corrupted speech signal is the sum of clean (original) speech signal and additive noise model. The degraded speech signal is processed by considering both low signal to noise ratio (SNR) and high SNR regions. The degraded speech segments are processed frame by frame with a duration of 20 ms. The SS-VAD method was proposed for speech enhancement in Refs. [1, 2, 4, 11, 17]. The effect of noise can be eliminated in degraded speech signal by subtracting the average magnitude spectrum of a noise model from the average magnitude spectrum of a degraded speech signal. The process of enhancing the degraded speech data using various noise elimination techniques is called speech processing [20]. The modified spectral subtraction algorithm was proposed for speech enhancement in Ref. [36]. This algorithm was implemented using VAD and minima-controlled recursive averaging (MCRA) [3]. The experimental results are evaluated under ITU Telecommunication Standardization Sector (ITU-T) G.160 standard and compared with existing methods. In Ref. [18], an improved spectral subtraction algorithm was proposed for musical noise suppression in degraded speech signal. The VAD was used for the detection of voiced regions in degraded speech signal. The experimental results show that there was more suppression of musical noise in degraded speech signal.

Various speech signal magnitude squared spectrum (MSS) estimators were proposed for noise reduction in degraded speech signal [8, 9, 22]. The MSS estimators, namely, minimum mean square error-short time power spectrum (MMSE-SP), minimum mean square error-spectrum power based on zero crossing (MMSE-SPZC), and maximum a posteriori (MAP) are implemented individually. These MSS estimators significantly performed well under many degraded conditions [22]. Ephraim and Malah have proposed a minimum mean square error short time spectral amplitude (MMSE-STSA) estimator for speech enhancement [8]. This method was compared with most widely used algorithms such as spectral subtraction and Wiener filtering. It was observed that the proposed MMSE-STSA method gives better performance than the existing methods. An alternative to the Ephraim and Malah speech enhancement method was proposed under the assumption that the Fourier series expansion of clean (original) speech signal and noise may be modeled independently with zero mean and Gaussian random variables [35]. Rainer Martin proposed an algorithm for speech enhancement using MMSE estimators and super-Gaussian priors [24]. The main advantage of this algorithm was to improve the short time spectral coefficients of corrupted speech signal. This method was compared with Wiener filtering and MMSE-STSA methods [8]. Philipos C. Loizou proposed an algorithm for noise reduction in corrupted speech signal using Bayesian estimators [19]. Three different types of Bayesian estimators are implemented for speech enhancement.

A few years back, two advanced acoustic modeling approaches, namely, subspace Gaussian mixture model (SGMM) and deep neural network (DNN) were described in Refs. [5, 12, 13, 26]. These two techniques

provide better speech recognition performance than the GMM-HMM-based approach. The more compact representation of GMM in acoustic space is SGMM. Therefore, the SGMM is best suitable for the moderate training speech data. Furthermore, the DNN consists of multiple hidden layers in multilayer perception to capture the nonlinearities of training set. This gives a good improvement in the modeling of the variations of acoustics leading to a better performance of speech recognition. The main contributions are made in this work are as follows:

- Deriving and studying the effectiveness of existing and newly proposed noise elimination techniques for practically deployable ASR system.
- The size of the speech database is increased by collecting 50 h of farmers' speech data (2000 farmers' speech data were collected in earlier work [32]) from different dialect regions of Karnataka and creating an entire dictionary (including all districts, mandis, and commodities of Karnataka state as per AGMARKNET list) and phoneme set for the Kannada language.
- Exploring the efficacy of the SGMM and DNN for moderate ASR vocabulary.
- Improving the on-line and off-line (word error rates (WERs) of ASR models) speech recognition accuracy in Kannada spoken query system.
- Testing the newly developed spoken query system from farmers of Karnataka under uncontrolled environment.

The remainder of the paper is organized as follows: Section II describes the speech data collection and preparation. The background noise elimination by combining SS-VAD and MMSE-SPZC estimator is described in Section III. The development of ASR models using Kaldi is described in Section IV. The effectiveness of SGMM and DNN is described in detail in Section V. The experimental results and analysis are discussed in Section VI. Section VII gives the conclusions.

2 Farmers' Speech Data Collection

The basic building block diagram of the different steps involved in the development of the improved Kannada spoken query system is shown in Figure 1.

An Asterisk server (Digium, Huntsville, AL, USA) is used in this work, which acts as an interface between the user/farmer, and the IVRS call flow is shown in Figure 2. Along with Asterisk, the asterisk gateway interface (AGI) is an important module for system integration.

For the development of ASR models, the task-specific speech data were collected from 2000 farmers across the different dialect regions of Karnataka under degraded conditions [32]. In addition to this, another 500 farmers' speech data are collected to increase the training data set. The training and testing speech

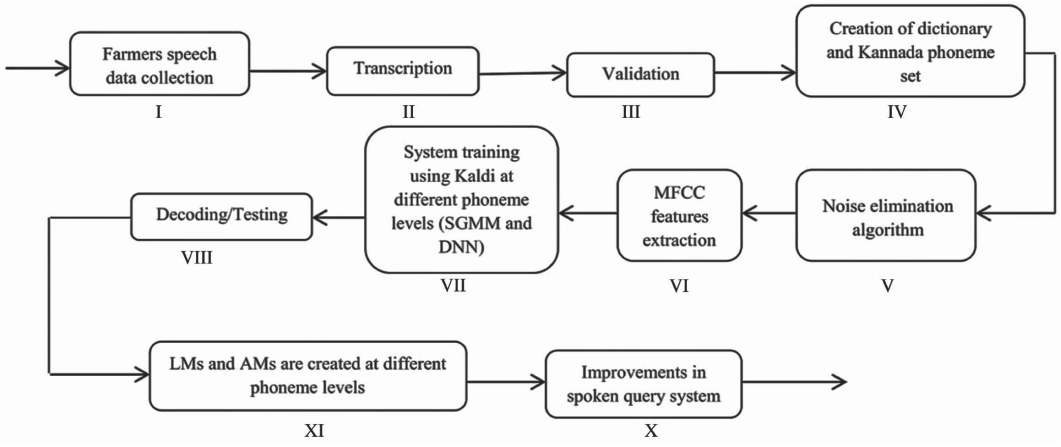


Figure 1: Block Diagram of Different Steps Involved in the Development of Kannada Spoken Query System.

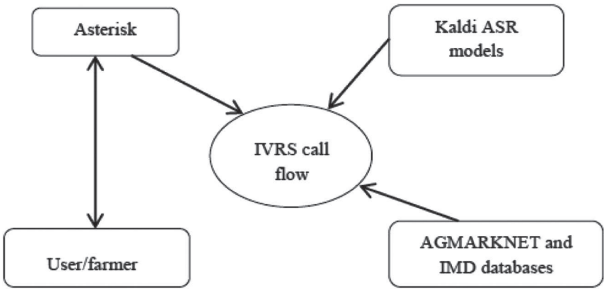


Figure 2: An Integration of ASR System and Asterisk.

data set consists of 70,757 and 2180 isolated word utterances, respectively. The training and testing data set includes the names of districts, mandis, and different types of commodities as per the AGMARKNET list under the Karnataka section. The performance estimation of the entire ASR system is done by overall speech data, which includes districts, mandis, and commodities.

3 Combined SS-VAD and MMSE-SPZC Estimator for Speech Enhancement

A noise elimination technique is proposed for speech enhancement, which is a combination of SS-VAD and MMSE-SPZC estimator. Consider a clean (original) speech signal s(n), which is corrupted by background noise d(n) that leads to the corrupted speech signal c(n). It can be written as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

The above assumption is usually used in traditional spectral subtraction algorithms. In the remainder of the paper, we will be calling C_k^2, S_k^2 and D_k^2 as magnitude squared spectrums of corrupted speech signal, clean speech signal, and noise model, respectively. The second assumption is that the complex part of the discrete Fourier transform (DFT) coefficients is modeled as free Gaussian random variables.

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

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

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

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

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

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

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

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

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

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

and

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

Note: If $\sigma_s^2(k) > \sigma_d^2(k)$, then $\frac{1}{\lambda(k)}$ is less than 0, and it is reversible. Hence, Ψ_k in equation (12) is positive.

3.1 Spectral Subtraction with VAD

The spectral subtraction method is commonly used for noise cancellation in degraded speech signal. The VAD plays an important role in the detection of only voiced area in the speech signal. In this method, we have considered the clean speech signal with 6-s duration, and it is corrupted by different noises, including additive white Gaussian noise (AWGN), car, factory, and f16 noises. The corrupted speech signal c(n) is converted

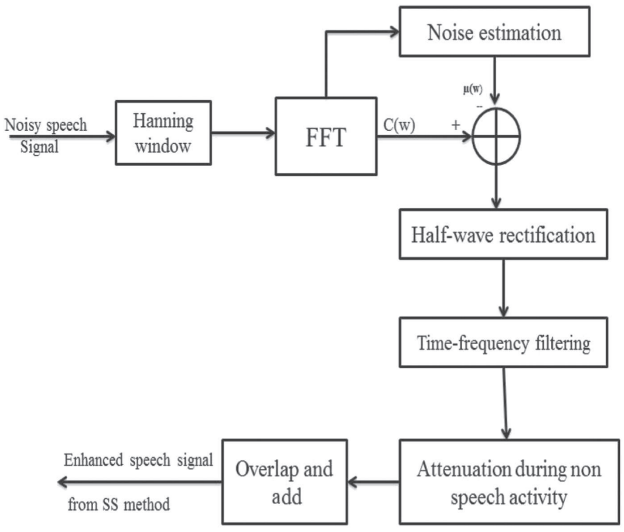


Figure 3: Block Diagram of SS-VAD Method.

into segments, and each segment consists of 256 samples with the sampling frequency of 8 kHz. The frame overlapping rate of 50% is considered, and the Hanning window is used in this work.

The basic building block diagram of SS-VAD is given in Figure 3. It consist of several main steps, namely, windowing, fast Fourier transform (FFT) calculation, noise estimation, half wave rectification, residual noise reduction, and calculation of inverse fast Fourier transform (IFFT). The corrupted speech signal c(n) is the combination of clean (original) speech signal s(n) and additive background noise d(n). The corrupted speech signal c(n) is given as an input to the spectral subtractor. The corrupted speech signal is Hanning windowed, and the FFT is calculated. The FFT is one of the most important methods to analyze the speech spectrum. The active regions of speech signal are identified by VAD; hence, the noise is estimated. The linear prediction error (LPE) is mainly associated with energy E of the signal and zero crossing rate (ZCR). The parameter Y can be written as follows:

$$Y = E(1 - Z)(1 - E) \text{ for single frame} \tag{15}$$

$$Y_{\max} = Y \text{ for all frames} \tag{16}$$

where Z and L are ZCR and LPE, respectively. The fraction term $\frac{Y}{Y_{\max}}$ is used to know whether a signal has voice activity or not. The average magnitude spectrum of the VAD output is subtracted with the average magnitude spectrum of the noise estimated. Hence, this process is called spectral subtraction with VAD. The output of spectral subtraction with VAD can be written as follows:

$$|X_i(w)| = |C_i(w)| - |\mu_i(w)| \tag{17}$$

where $w = 0, 1, 2, \ldots, L - 1$ and $i = 0, 1, 2, \ldots, M - 1$. The term L indicates the length of FFT, and M indicates the number of frames. The half wave rectifier is used in this work to set the spectrums' negative values to zero if they have negative values. Reduction of residual noise in enhanced speech signal is the final step of spectral subtraction. During the non-speech activities, it is needed to further attenuate the signal. This improves the quality of the enhanced speech signal. Finally, the enhanced speech signal is obtained by calculating its IFFT. The enhanced speech data could be used in speech processing applications, such as speech recognition, speaker identification, speaker verification, and speaker recognition, etc.

3.2 Magnitude Squared Spectrum Estimators

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

– Minimum mean square error-short-time power spectrum (MMSE-SP) estimator.
– Minimum mean square error-spectrum power estimator based on zero crossing (MMSE-SPZC).
– Maximum a posteriori (MAP) estimator.

3.2.1 Minimum Mean Square Error-Short Time Power spectrum (MMSE-SP) Estimator

In Ref. [35], the authors proposed an algorithm for the MMSE-SP estimator. The clean (original) speech signal can be obtained by taking the expected value of the clean speech signal and Fourier transform of the corrupted speech signal C(w). It can be written as follows:

$$\hat{S}_K^2 = E\{S_k^2|C(w_k)\} \tag{18}$$

$$\hat{S}_K^2 = \int_0^\infty S_k^2 f_{S_k}(S_k|C(w_k)\ dS_k \tag{19}$$

$$\hat{S}_K^2 = \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k}+\frac{\xi_k}{1+\xi_k}\right)C_k^2 \tag{20}$$

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

$$\xi_k \equiv \frac{\sigma_s^2(k)}{\sigma_d^2(k)},\ \gamma_k \equiv \frac{C_k^2}{\sigma_d^2(k)} \tag{21}$$

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

$$f_{S_k}(S_k|C(w_k)) = \frac{S_k}{\sigma_k^2} exp\left(\frac{S_k^2+u_k^2}{2\sigma_k^2}\right) I_0\left(\frac{S_k u_k}{\sigma_k^2}\right) \tag{22}$$

where

$$\frac{1}{\lambda'(k)} \equiv \frac{1}{\sigma_s^2(k)}+\frac{1}{\sigma_d^2(k)} \tag{23}$$

$$v_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k \tag{24}$$

$$\sigma_k^2 \equiv \frac{\lambda'(k)}{2} \text{ and } u_k^2 \equiv v_k\lambda'(k) \tag{25}$$

where $I_0(\cdot)$ is the 0^{th}-order modified Bessel function. The approximate values of the Bessel function are calculated in order to derive magnitude spectrums of the MAP estimator [21].

3.2.2 Minimum Mean Square Error-Spectrum Power Estimator Based on Zero Crossing (MMSE-SPZC)

Another important magnitude squared spectrum estimator is the MMSE-SPZC. Using the work, which was presented in Refs. [1] and [4], the MMSE-SPZC estimator is derived [22]. The initial MMSE estimator is obtained

by calculating the mean of the a posteriori density function as shown in equation (7).

$$\hat{S}_k^2 = E\{S_K^2|C_k^2\} \tag{26}$$

$$\hat{S}_K^2 = \int_0^{C_k^2} S_k^2 f_{S_k^2}(S_k^2|C_k^2) dS_k^2 \tag{27}$$

$$\hat{S}_K^2 = \begin{cases} \left(\frac{1}{\nu_k} - \frac{1}{e_k^{\nu} - 1}\right) C_k^2, & \text{if } \sigma_s^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} C_k^2, & \text{if } \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{28}$$

where ν_k can be written as

$$\nu_k \equiv \frac{1 - \xi_k}{\xi_k} \gamma_k \tag{29}$$

The estimator gain function can be represented mathematically as follows:

$$G_{MMSE}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{\nu_k} - \frac{1}{e_k^{\nu} - 1}\right)^{\frac{1}{2}} & \text{if } \sigma_s^2(k) \neq \sigma_d^2(k) \\ \left(\frac{1}{2}\right)^{\frac{1}{2}} & \text{if } \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{30}$$

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

3.2.3 Maximum A Posteriori (MAP) Estimator

The MAP estimator can be represented as follows:

$$\hat{S}_k^2 = argmax\, f_{S_K^2}(S_k^2|C_k^2) \tag{31}$$

Maximization with respect to S_K^2.

$$\hat{S}_k^2 = \begin{cases} C_k^2 & \text{if } \frac{1}{\lambda(k)} < 0 \\ 0 & \text{if } \frac{1}{\lambda(k)} > 0 \end{cases} \tag{32}$$

$$\hat{S}_k^2 = \begin{cases} C_k^2 & \text{if } \sigma_s^2(k) \geq \sigma_d^2(k) \\ 0 & \text{if } \sigma_s^2(k) < \sigma_d^2(k) \end{cases} \tag{33}$$

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

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

$$G_{MAP}(k) = \begin{cases} 1 & \text{if } \sigma_s^2(k) \geq \sigma_d^2(k) \\ 0 & \text{if } \sigma_s^2(k) < \sigma_d^2(k) \end{cases} \tag{34}$$

Using equation (22), the above MAPs gain function can also be represented as:

$$G_{MAP}(\xi_k) = \begin{cases} 1 & \text{if } \xi_k \geq 1 \\ 0 & \text{if } \xi_k < 1 \end{cases} \tag{35}$$

It can be observed in the above equation that the gain function of the MAP estimator is binary in nature. In fact, it is almost the same as the binary mask, which is widely used in computational auditory scene analysis (CASA) [34]. The gain function of the MAP estimator is based on an a priori SNR, and the gain function of the binary mask is based on instantaneous SNR, and this makes a difference between them. The MAP estimator uses the hard thresholding algorithm, which can be most widely used in wavelet shrinkage algorithm [6, 7, 16, 23].

3.3 Performance Measures and Analysis

The performance of existing methods and proposed methods is evaluated from the standard measures. They are perceptual evaluation of speech quality (PESQ) and composite measure described below.

3.3.1 PESQ

The PESQ measure is an objective measure and it is strongly recommended by ITU-T for quality of speech assessment [15, 30]. The term PESQ is calculated as the linear sum of the average distortion value D_{ind} and average asymmetrical distortion value A_{ind}. It can be written as follows [25]:

$$PESQ = b_0 + b_1 D_{ind} + b_2 A_{ind} \tag{36}$$

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

3.3.2 Composite Measures

Composite measures are the objective measures, which can be used for the performance evaluation. The ratings and description of the different scales are shown in Table 1. The composite measures are derived by multiple linear regression analysis [14]. The multiple linear regression analysis is used to estimate the three important composite measures [15]. They are:

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

3.4 Performance Analysis of Existing Methods

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

Table 1: The Description of the Speech Signal Distortion (s), Background Noise Distortion (b), and Overall Speech Quality (o) Scales Rating.

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

3.4.1 Spectral subtraction With VAD Results and Analysis

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

3.4.2 Magnitude Squared Spectrum Estimator Results and Analysis

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

Table 2: Performance Measurement of SS-VAD Method in Terms of PESQ for TIMIT Database.

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

Table 3: Performance Measurement of SS-VAD Method in Terms of PESQ for Kannada Database.

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

Table 4: Performance Evaluation of SS-VAD Method Using Composite Measure for TIMIT Database.

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

Table 5: Performance Evaluation of SS-VAD Method Using Composite Measure for Kannada Database.

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

Table 6: Performance Measurement of MSS Estimators in Terms of PESQ for TIMIT Database.

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

Table 7: Performance Measurement of MSS Estimators in Terms of PESQ for Kannada Database.

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

Table 8: Performance Evaluation of MSS Estimators Using Composite Measure for TIMIT Database.

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

Table 9: Performance Evaluation of MSS Estimators Using Composite Measure for Kannada Database.

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

3.5 Proposed Combined SS-VAD and MMSE-SPZC Method

The SS-VAD method suppresses the various types of noises reasonably such as babble noise, street noise, car noise, vocal noise, and background noise, etc. The main drawback of the SS-VAD is that the suppression of musical noise in degraded speech signal is much less [2, 4, 20]. The MMSE-SPZC is a robust method to suppress the musical noise given the better results for car, street, and white noises compared to babble noise [22]. Therefore, to overcome the problem of suppression of musical and babble noises, a method is proposed. The proposed method is a combination of the above two methods, which suppresses the different types of noises including musical and babble noise reasonably under uncontrolled environment. The flowchart of the proposed method is shown in Figure 4. The output of the SS-VAD is a little noisier, and musical noise is not suppressed as well. Therefore, the output of the SS-VAD is passed through the MMSE-SPZC estimator.

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

$$|X_i(w)| = |C_i(w)| - |\mu_i(w)| \tag{37}$$

where $w = 0, 1, 2, \ldots, L-1$ and $i = 0, 1, 2, \ldots, M-1$. The MMSE-SPZC estimator is derived once again for the SS-VAD output. The output x(n) is passed through the MMSE-SPZC estimator. Hence, the MMSE-SPZC estimator is derived by considering the mean of the a posteriori density function of the SS-VAD output.

$$\hat{X}_k^2 = E\{X_K^2 | Y_k^2\} \tag{38}$$

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

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

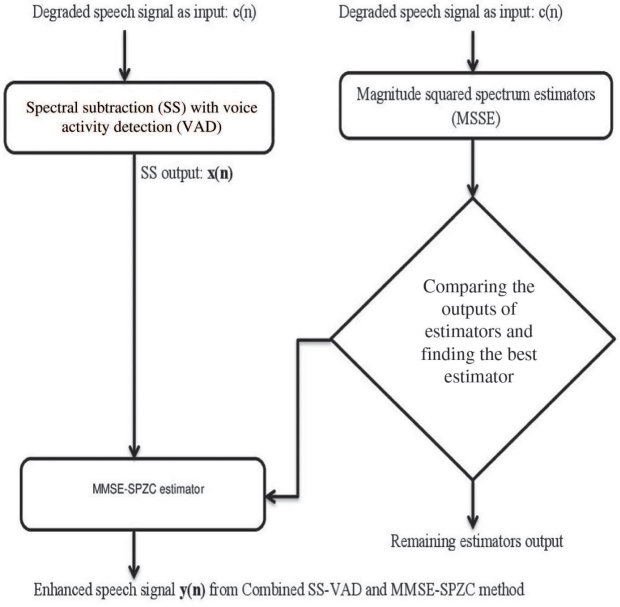


Figure 4: Flow Chart of Combined SS-VAD and MMSE-SPZC Method.

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

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

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

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

The description of the performance measurement of the proposed method in terms of the PESQ for both the databases is shown in Tables 10 and 11. From the tables, it was observed that there is much suppression in the babble and musical noises with PESQ improvements of 0.6933, 0.7781 and 0.7112, 0.7314 for the TIMIT and Kannada databases, respectively, by the proposed method compared to the individual methods. The speech quality is much improved after the performance evaluation of the proposed method using composite measures for both the databases and is shown in Tables 12 and 13. From the experimental results and analysis, it can be inferred that the combined SS-VAD and MMSE-SPZC method reduces the noise in the degraded speech data significantly compared to the individual methods. The enhanced speech data obtained from the proposed method is more audible and has a higher quality than the individual methods. Therefore, the proposed method can be used for the Kannada speech database enhancement as the majority of the collected speech data is degraded by babble, musical, and street noises as it is collected under an uncontrolled environment. When the proposed noise elimination technique is applied to both training and testing speech data sets, it is found to significantly improve the performance (off-line and on-line) of the spoken query system.

Table 10: Performance Measurement of Combined SS-VAD and MMSE-SPZC Estimator in Terms of PESQ for TIMIT Database.

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

Table 11: Performance Measurement of Combined SS-VAD and MMSE-SPZC Estimator in Terms of PESQ for Kannada Database.

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

Table 12: Performance Evaluation of Combined SS-VAD and MMSE-SPZC Method Using Composite Measure for TIMIT Database.

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

Table 13: Performance Evaluation of Combined SS-VAD and MMSE-SPZC Method Using Composite Measure for Kannada Database.

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

4 Creation of ASR Models Using Kaldi for Noisy and Enhanced Speech Data

The ASR models (language and acoustic models) play an important role in the development of robust speech recognition systems. The language models (LMs) and acoustic models (AMs) were developed in Ref. [32] for the noisy speech data collected in the field. Therefore, the ASR models are developed in this work for the noisy and enhanced speech data. The development of the ASR models includes several steps. They are as follows:

- Transcription and validation of enhanced speech data.
- Development of lexicon (dictionary) and Kannada phoneme set.
- MFCC feature extraction.

4.1 Transcription and Validation

Transcription is a process of converting speech file content into its word format and its equivalent, which is also called as a word-level transcription. The schematic representation of various speech wave files and those equivalent transcriptions are shown in Figure 5.

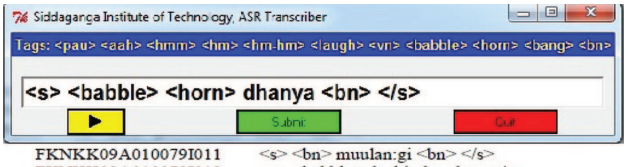
It was observed in the figure that the tags <s> and </s> indicate the starting and ending of speech sentence/utterance. The different types of tags are used in the transcription. They are as follows:

- <music>: Used only when the speech file is degraded by music noise.
- <babble>: Used only when the speech file is degraded by babble noise.
- <bn>: Used only when the speech file is degraded by background noise.
- <street>: Used only when the speech file is degraded by street noise.

If the transcription of a particular speech datum is done wrongly, then, it will be validated using the validation tool shown in Figure 6. The speech file *dhanya* is degraded by background noise, babble noise, horn noise, and musical noise, but unknowingly, the transcriber transcribed the same speech file into <**s**> <**babble**> <**horn**> **dhanya** <**bn**> <**/s**> only. While cross checking the transcribed speech datum, the validator listened to that speech sound again and found that it is degraded by musical noise also. Therefore, it could be validated as <**s**> <**babble**> <**horn**> **dhanya** <**bn**> <**music**> <**/s**> and is shown in Figure 6.

4.2 Kannada Phoneme Set and Corresponding Dictionary Creation

The Karnataka is one of the states in India. There are 60 million people living in Karnataka who fluently speak the Kannada language. The Kannada language has 49 phonetic symbols, and it is one of the most



FKNKK09A010079I011 <s> <bn> muulan:gi <bn> </s>
FKNKK09A010079I012 <s> <babble> daal:in:be <bn> </s>
FKNKK09A010079I013 <s> <horn> ged:d:ekoosu <street> <bn> </s>
FKNKK09A010079I014 <s> <babble> <horn> dhanya <bn> </s>
FKNKK09A010079I015 <s> <bn> bat:aand~i <bang> </s>
FKNKK09A010079I016 <s> <bang> avarekaayi <bn> </s>
FKNKK09A010079I017 <s> nuggeikaayi <bn> </s>
FKNKK09A010079I018 <s> <street> bella <bn> <bang> </s>
FKNKK09A010079I019 <s> <bang> seibu <bang> </s>
FKNKK09A010079I020 <s> <bn> mekkejool:a <vn> <bn> </s>
FKNKK09A010079I021 <s> <bn> uddu <babble> <bn> </s>
FKNKK09A010079I022 <s> <bn> mend~asu <bn> </s>
FKNKK09A010079I023 <s> <bn> an:t:avaal:a <bn> </s>
FKNKK09A010079I024 <s> <bn> <bang> good:an:bi <bn> </s>
FKNKK09A010079I025 <s> <bn> ten:gu </s>
FKNKK09A010079I026 <s> <bn> <vn> navand~e </s>

Figure 5: Transcription of Speech Data.

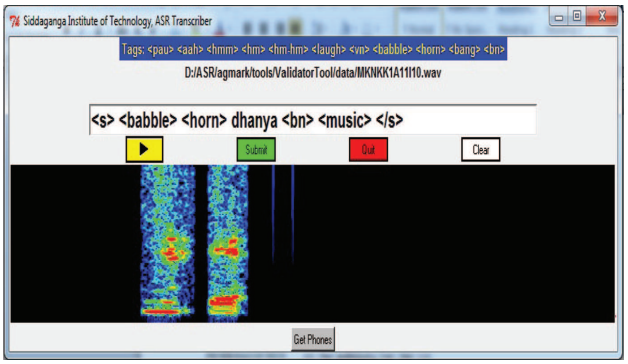


Figure 6: Validation Tool to Validate the Transcribed Speech Data.

Table 14: The Labels Used from the Indic Language Transliteration Tool (IT3 to UTF-8) for Kannada Phonemes.

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

Table 15: The Labels Used from ILSL12 for Kannada Phonemes.

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

usable Dravidian languages. The description of the Kannada phoneme set, Indian language speech sound label 12 (ILSL12) set used for Kannada phonemes, and its corresponding dictionary are shown in Tables 14–16, respectively.

4.3 MFCC Feature Extraction

Once the noise elimination algorithm is applied on the train and test data set, the next step is to extract the MFCC features for noisy and enhanced speech data. The basic building block diagram of the MFCC feature extraction is shown in Figure 7.

The parameters used for the MFCC feature extraction are as follows:

- Window used: Hamming window.
- Window length: 20 ms.

Table 16: Dictionary/Lexicon for Some of Districts, Mandis, and Commodities Enlisted in AGMARKNET.

Label set using IT3-UTF:8	Label set using from ILSL12
daavand agere	d aa v a nx a g e r e
tumakuuru	t u m a k uu r u
chitradurga	c i t r a d u r g a
ben:gal:uuru	b e ng g a lx uu r u
chaamaraajanagara	c aa m a r aa j a n a g a r a
shivamogga	sh i v a m o gg a
haaveiri	h aa v ei r i
gadaga	g a d a g a
gulbarga	g u l b a r g a
hubbal:l:i	h u bb a llx i
dhaarawaad:a	d aa r a v aa dx a
bel:agaavi	b e lx a g aa v i
raamanagara	r aa m a n a g a r a
koolaara	k oo l aa r a
koppal:a	k o pp a lx a
raayacuuru	r aa y a c uu r u
chin:taamand i	c i n t aa m a nx i
tiirthahal:l:i	t ii r th a h llx i
harihara	h a r i h a r a
channagiri	c a nn a g i r i
honnaal:i	h o nn aa lx i
bhadraavati	bh a d r aa v a t i
theerthahal:l:i	t ii r th a h a llx i
kund igal	k u nx i g a l
tipat:uuru	t i p a tx uu r u
gubbi	g u bb i
korat:agere	k o r a tx a g e r e
akki	a kk i
raagi	r aa g i
jool:a	j oo lx a
mekkejool:a	m e kk e j oo lx a
bhatta	bh a tt a
goodhi	g oo dh i
iirul:l:i	ii r u llx i
ond amend asinakaayi	o nx a m e nx a s i n a k aa y i

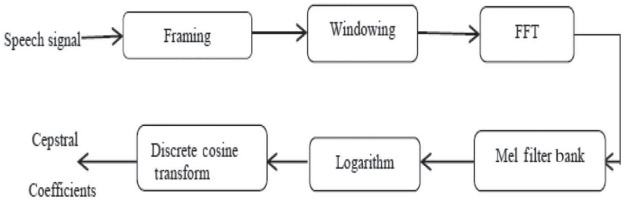


Figure 7: Block Diagram of MFCC Feature Extraction.

- Pre-emphasis factor: 0.97.
- MFCC coefficients: 13 dimensional.
- Filter bank used: 21-channel filter bank.

5 SGMM and DNN

The SGMM and DNN ASR modeling techniques are described in this section.

5.1 SGMM

The ASR systems based on the GMM-HMM structure usually involves completely training the individual GMM in every HMM state. A new modeling technique that is introduced to the speech recognition domain is called the SGMM [26]. The dedicated multivariate Gaussian mixtures are used for the state level modeling in conventional GMM-HMM acoustic modeling technique. Therefore, no parameters are distributed between the states. The states are represented by Gaussian mixtures, and these parameters distribute a usual structure between the states in the SGMM modeling technique. The SGMM consists of the GMM inside every context-dependent state; the vector $I_i \in V^r$ in every state is specified instead of defining the parameters directly.

An elementary form of the SGMM can be described by the equations below:

$$p(y|i) = \sum_{k=1}^{L} w_{ik} N(y; \mu_{ik}, \Sigma_k) \tag{42}$$

$$\mu_{ik} = M_k I_i \tag{43}$$

$$w_{ik} = \frac{expw_k^T I_i}{\sum_{k'=1}^{L} expw_{k'}^T I_i} \tag{44}$$

where y $\in$ $^{R^D}$ is a feature vector and $i \in \{1 \ldots I\}$ is the context-dependent state of speech signal. The speech state j's model is a GMM with L Gaussians (L is between 200 and 2000), with matrix of covariances Σ_k, which are distributed amidst states, mixture weights w_{ik}, and means μ_{ik}. The derivation of the $\mu_{ik} w_{ik}$ parameters are done using I_i together with M_k, w_k and Σ_k. The detailed description of parameterization of the SGMM and its impact is given in Ref. [31]. The ASR models are developed using this modeling technique for the Kannada speech database, and the least word error rate (WER) models could be used in the spoken query system.

5.2 DNN

The GMM-HMM-based acoustic modeling approach is inefficient to model the speech data that lie on or near the data space. The major drawbacks of the GMM-HMM-based acoustic modeling approach are discussed in Ref. [13]. The artificial neural networks (ANN) are capable of modeling the speech data that lie on or near the data space. It is found to be infeasible to train an ANN using the maximum number of hidden layers with back propagation algorithm for a large amount of speech data. An ANN with a single hidden layer failed to give good improvements over the GMM-HMM-based acoustic modeling technique. Both the aforementioned limitations were overcome with the developments in the past few years. Various approaches are available now to train the different neural nets with a maximum number of hidden layers.

The DNN consists of the maximum number of input hidden layers and output layers to model the speech data to build the ASR systems. The posterior probabilities of the tied states are modeled by training the DNN. This yielded a better performance in recognition compared to the conventional GMM-HMM acoustic modeling approach. The stacking layers of the restricted Boltzmann machine are used to create the DNN. The restricted Boltzmann machine is a undirected model and is shown in Figure 8.

The model uses the single parameter set (W) to state the joint probability variables vector (v) and hidden variables (h) through an energy E and can be written as follows:

$$p(v, h; W) = \frac{1}{Z} e^{-E(v,h;W)} \tag{45}$$

where Z is a function of partition, which can be written as

$$Z = \sum_{v',h'} e^{-E(v',h';W)} \tag{46}$$

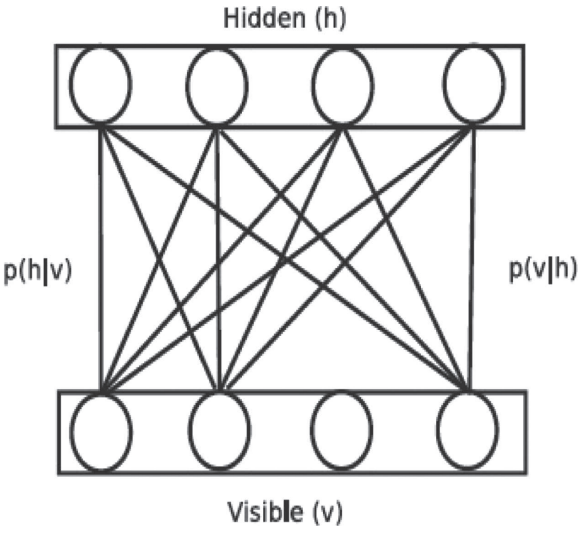


Figure 8: Block Diagram of Restricted Boltzmann Machine.

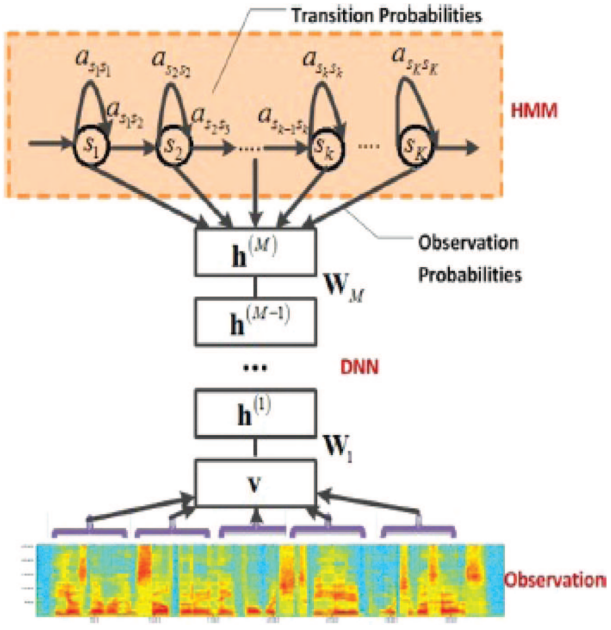


Figure 9: Block Diagram of Hybrid DNN and HMM.

where v' and h' are the extra variables used for the summation over the ranges of v and h. The unsupervised technique is described in detail in Ref. [12] for modeling the connection weights in deep belief networks that is approximately equal to training the next pair of restricted Boltzmann machine layers. The schematic representation of the context-dependent DNN-HMM hybrid architecture is shown in Figure 9. The modeling of tied states (senones) is done by context-dependent DNN-HMM. The MFCC features are given as input to the DNN input layer, and the output of the DNN is used with the HMM, which models the sequential property of the speech.

6 Experimental Results and Analysis

The system training and testing using Kaldi is done in two phases, one for noisy speech data and another for enhanced speech data. The 90% and 10% of the validated speech data is used for training and testing, respectively. The number of speech files used for system training and testing is shown in Table 17.

Table 17: The Speech Files Used for System Training and Testing.

Kannada speech data	Number of train files	Number of test files
Overall noisy speech data	68523	2180
Overall enhanced speech data	67234	2180

Table 18: The Description of WERs at Different Phoneme Levels for Overall Noisy Speech Data, Which Includes Districts, Mandis, and Commodities.

Phoneme level	WER 1	WER 2	WER 3	WER 4	WER 5	WER 6
mono	31.61	31.63	31.88	31.73	31.62	31.63
tri1_2000_8000	16.15	16.16	16.18	16.21	16.23	16.19
tri1_2000_16000	14.95	14.99	14.98	14.98	15.01	15.11
tri1_2000_32000	14.63	14.71	14.78	14.69	14.71	14.77
tri2	13.35	13.36	13.39	13.39	13.37	13.41
tri2_2000_8000	15.14	15.18	15.20	15.19	15.21	15.13
tri2_2000_16000	13.85	13.89	13.88	13.90	13.87	13.86
tri2_2000_32000	13.12	13.19	13.20	13.11	13.09	13.21
tri3_2000_8000	14.91	14.90	14.92	14.97	14.96	14.95
tri3_2000_16000	13.78	13.76	13.76	13.71	13.79	13.78
tri3_2000_32000	13.17	13.20	13.21	13.23	13.29	13.30
Sgmm	12.98	12.99	12.86	**12.78**	12.87	12.79
tri4_nnet_t2a	11.88	11.89	11.89	11.86	11.88	**11.82**

The LMs and AMs are developed for 50 h of speech data. A total of 70,757 isolated speech utterances are used for overall noisy speech data training and testing. In these, 68,523 utterances are used for system training, and 2180 utterances are used for testing to build ASR models for overall noisy speech data. Likewise, 69,268 utterances are used for overall enhanced speech data training and testing using Kaldi. In those, 67,234 utterances used for system training and 2180 utterances are used for testing to build the ASR models for the overall enhanced speech data shown in Table 17. In this work, 62 non-silence phones, nine silence phones are used, and "sil" is used as the optional silence phone. The LMs and AMs are created at different phoneme levels and are as follows:

- Monophone training and decoding.
- Triphone1: Deltas + Delta-Deltas training and decoding.
- Triphone2: linguistic data analysis (LDA) + maximum likelihood linear transform (MLLT) training and decoding.
- Triphone3: LDA + MLLT + speaker adaptive training (SAT) and decoding.
- SGMM training and decoding.
- DNN hybrid training and decoding.

The 2000 senons and 4, 8, and 16 Gaussians mixtures are used in this work to build the ASR models at monophone, triphone1, triphone2, and triphone3 levels. The recently introduced two modeling techniques, such as the SGMM DNN, are used to build the ASR models. The off-line speech recognition performance is measured by word error rate (WER). Table 18 shows the description of WERs at different phoneme levels for overall noisy speech data (combined districts, mandis, and commodities). The WERs of 12.78% and 11.82% are achieved for the SGMM and hybrid DNN training and decoding, respectively, for overall noisy speech data.

The WERs of 11.77% and 10.67% are obtained for the overall enhanced speech data using the SGMM training and decoding and hybrid DNN training and decoding, respectively, as shown in Table 19.

6.1 Call Flow Structure of Spoken Query System

The ASR models were developed in the earlier work [32] at monophone, triphone1, and triphone2 levels with 600 senons and 4, 8, and 16 Gaussian mixtures. The achieved least WERs are 10.05%, 11.90%, 18.40%, and

Table 19: The Description of WERs at Different Phoneme Levels for Overall Enhanced Speech Data, Which Includes Districts, Mandis, and Commodities.

Phoneme level	WER 1	WER 2	WER 3	WER 4	WER 5	WER 6
mono	30.55	30.57	30.58	30.55	30.56	30.59
tri1_2000_8000	15.52	15.50	15.58	115.60	15.61	15.50
tri1_2000_16000	13.90	13.89	13.92	13.94	13.99	13.91
tri1_2000_32000	13.48	13.50	13.48	13.49	13.51	13.52
tri2	12.55	12.52	12.53	12.55	12.58	12.54
tri2_2000_8000	14.79	14.77	14.78	14.80	14.81	14.79
tri2_2000_16000	12.93	12.91	12.94	12.99	13.00	13.01
tri2_2000_32000	12.29	12.30	12.31	12.30	12.34	12.37
tri3_2000_8000	13.99	13.98	14.00	14.01	14.04	13.97
tri3_2000_16000	12.61	12.60	12.63	12.65	12.68	12.62
tri3_2000_32000	12.01	12.00	12.09	12.05	12.03	12.04
Sgmm	11.78	11.79	**11.77**	11.80	11.81	11.80
tri4_nnet_t2a	**10.67**	19.69	10.70	10.70	10.71	10.73

17.28% for districts, mandis, commodities, and overall speech data, respectively. This leads to the less recognition of commodities and mandis due to the high WERs. To overcome this problem, another 500 farmers' speech data are collected, the training data set is increased, and the noise elimination algorithm on both training and testing data set to improve the accuracy of the ASR models in this work is applied. In the earlier work, separate spoken query systems were developed to access the real-time agricultural commodity prices and weather information in the Kannada language [32]. In this work, the two call flow structures have been integrated together and made as a single call flow to overcome the complexity of the dialing multiple call flows. Therefore, the user/farmer can access both information in a single call flow by dialing a toll-free number. The earlier call flow structure of spoken query systems for agricultural commodity prices and weather information access are shown in Figures 10 and 11, respectively.

The schematic representation of the new integrated call flow structure is shown in Figure 12. Comparing the WERs of earlier work, a significant improvement in accuracy of 4% for mono, tri1, and triphone2 levels (for overall noisy speech data) is achieved. The ASR models are developed using tri3 (LDA+MLLT_SAT), SGMM, and DNN for overall noisy and enhanced speech data in this work. From the above tables, it can be observed that there is significant improvement in accuracy with less WER for enhanced speech data. Approximately, 1.2% of accuracy is improved for the speech data after speech enhancement. The ASR models are developed for overall speech data to reduce the complexity in the call flow of the spoken query system. The overall speech data include all districts, mandis, and commodities enlisted in the AGMARKNET website under the Karnataka section. The developed least WER models (overall enhanced speech data models) could be used in the spoken query system to improve its on-line recognition performance. The Kannada spoken query system needs to recognize 250 commodities including all its varieties. The three districts, 45 mandis, and 98 commodities are included in this work. The developed spoken query system enables the user/farmer to call the system. In the first step, the farmer needs to prompt the district name. If the district is recognized, then, the system asks for the mandi name. If the mandi name is also recognized, then, it will ask the farmer to prompt the commodity name. If the commodity name is recognized, then, it will play out the current price information of the asked commodity from the price information database. Similarly, to get weather information, the farmer needs a prompt, the district name, at the first step. If the district is recognized, then the system gives the current weather information through prerecorded prompts from the weather information database. If the district, mandi, and commodities are not recognized, then, the system gives two more chances to prompt those again. Nevertheless, these are not recognized properly, then, the system says, "Sorry. Try again later."

6.2 Testing of Developed Spoken Query System from Farmers in the Field

The spoken query system is again developed for the new ASR models. To check the on-line speech recognition accuracy of the newly developed spoken query system, the 300 farmers are asked to test the system

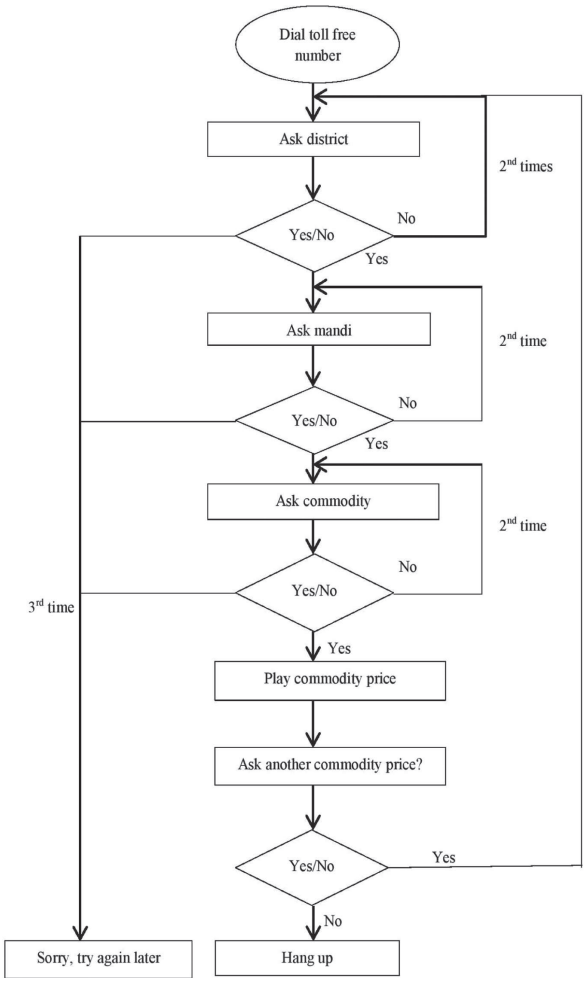


Figure 10: Call Flow Structure of Commodity Price Information Spoken Query System (Reproduced from Ref. [32] for Correct Flow of this Work).

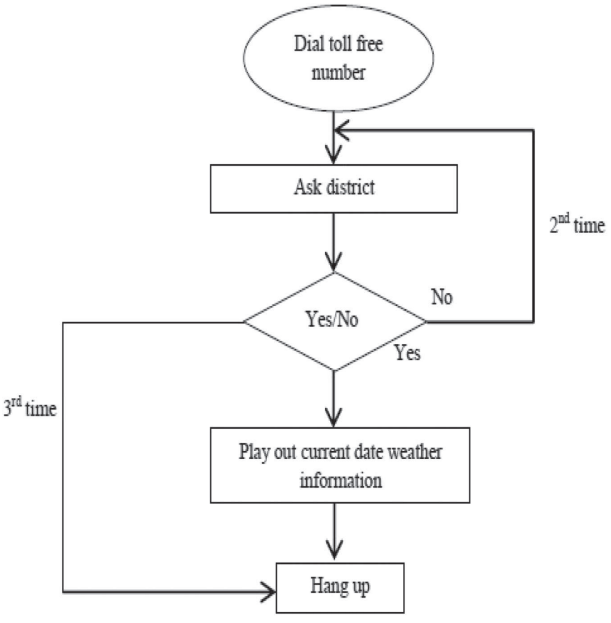


Figure 11: Call Flow Structure of Weather Information Spoken Query System (Reproduced from Ref. [32] for Completeness.

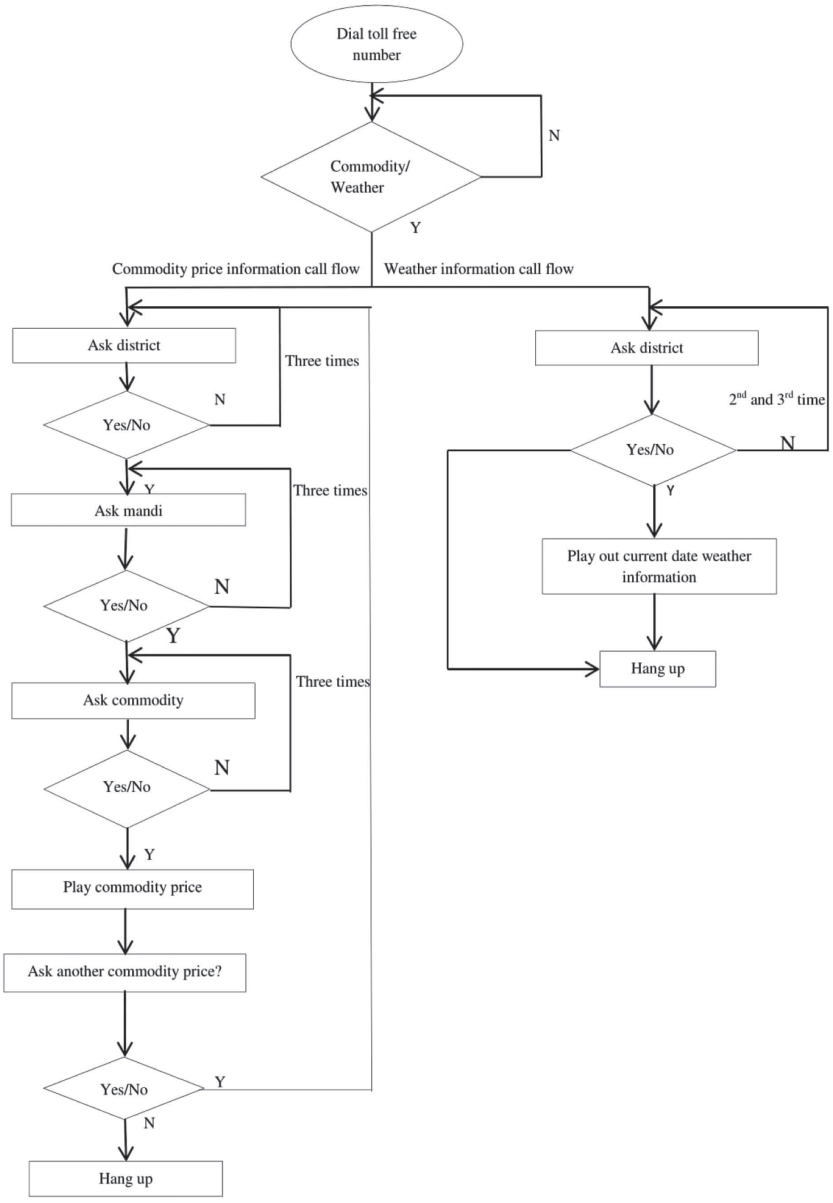


Figure 12: Call Flow Structure of Spoken Query System to Access Both Agricultural Commodity Prices and Weather Information.

Table 20: Performance Evaluation of On-Line Speech Recognition Accuracy Testing by Farmers in Field Conditions.

Language: Kannada	Total no. of farmers	1st attempt	2nd attempt	3rd attempt	Total no. of recognitions	Recognition in %
Districts	300	241	15	10	266	88.66
Mandis	300	238	18	10	262	87.33
Commodities	300	240	14	9	263	87.66

under uncontrolled environment. Table 20 shows the performance evaluation of the newly developed spoken query system by the farmers. It was observed in the table that there is a much improvement in on-line speech recognition accuracy with less failure of recognizing the speech utterances compared to the earlier spoken query system. Therefore, it can be inferred that the on-line and off-line (WERs of models) recognition rates are almost the same as shown in Tables 19 and 20.

7 Conclusions

The development and testing of the new Kannada spoken query system is demonstrated in this work. A method is proposed for the background noise mitigation, which is a combination of the SS-VAD and MMSE-SPZC estimator. The collected task-specific speech data was much degraded by musical, background, street, and babble noises. The proposed method gave better results compared to the individual methods for the TIMIT and Kannada speech databases. Therefore, it was applied on collected speech data for speech enhancement (before MFCC feature extraction). The 90% and 10% of validated speech data were used for system training and testing, respectively, using the Kaldi speech recognition toolkit. We also demonstrated the effectiveness of the recently introduced speech recognition modeling techniques SGMM and DNN. The ASR models were created for both noisy and enhanced speech data. The GMM-HMM-based acoustic modeling technique was used in the earlier spoken query system. The SGMM and DNN modeling approaches replaced the GMM-HMM-based acoustic modeling technique in this work. The SGMM-DNN ASR models are found to be outperformed compared to the conventional GMM-HMM-based acoustic models. There is a significant improvement in accuracy of 4% in the DNN-SGMM-based acoustic models compared to the earlier GMM-HMM-based models. Using the Kaldi recipe and Kannada language resources, the achieved WERs are 12.78% and 11.82% for the noisy speech data using the SGMM and hybrid DNN-based modeling techniques. The WERs of 11.77% and 10.67% are achieved for enhanced speech data using the same modeling techniques. Therefore, it can be inferred that there is a significant improvement in accuracy of 1.2% for the enhanced speech data compared to the noisy speech data. Interestingly, both the SGMM and DNN-based modeling approaches result in very similar WERs. The least WER models (SGMM and DNN-based models) could be used in the newly designed spoken query system. The earlier two spoken query systems are integrated together to form a single spoken query system. Therefore, the user/farmer can access the agricultural commodity price/weather information in a single call flow. The on-line speech recognition testing of the newly developed spoken query system done by farmers under real-time environments is also presented in this work. The future challenging work is to increase the Kannada speech database, further improving the performance of the ASR models, and testing the newly developed spoken query system by the farmers from the different dialect regions of the Karnataka state.

Acknowledgment: This work is a part of the ongoing consortium project on Speech Based Access of Agricultural Commodity Prices and Weather Information in 11 Indian Languages/Dialects funded by the Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology (MC&IT), Government of India. The authors would like to thank the consortium leader Prof. S. Umesh and other consortium members for their valuable inputs and suggestions.

Bibliography

[1] J. Beh and H. Ko, A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech, in: *IEEE Int. Conf. on Multimedia and Expo,* vol. 3, I-648, I-651, April 2003.

[2] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust. Speech Signal Process* **2** ASSP-27 (1979), 113–120.

[3] I. Cohen and B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement, *IEEE Signal Process. Lett.* **9** (2002), 12–15.

[4] C. Cole, M. Karam and H. Aglan, Spectral subtraction of noise in speech processing applications, in: *40th Southeastern Symposium System Theory, SSST-2008,* pp. 50–53, 16–18, New Orelans, LO, USA, March 2008.

[5] G. Dahl, D. Yu, L. Deng and A. Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition, in: *IEEE Trans. on Audio Speech, and Language Processing* (receiving 2013 IEEE SPS Best Paper Award), pp. 30–42, Piscataway, NJ, USA, 2012.

[6] D. L. Donoho and I. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81** (1994), 425–455.

[7] D. L. Donoho and I. M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *J. Am. Stat. Assoc.* **90** (1995), 1200–1224.

[8] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean square error short-time spectral amplitude estimator, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32** (1984), 1109–1121.

[9] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean square error log-spectral amplitude estimator, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-33** (1985), 443–445.
[10] J. R. Glass, Challanges for spoken dialogue systems, in: *Proc. IEEE ASRU Workshop*, Piscataway, NJ, USA, 1999.
[11] H. M. Goodarzi and S. Seyedtabaii, Speech enhancement using spectral subtraction based on a modified noise minimum statistics estimation, in: *Fifth Joint Int. Conf,* pp. 1339, 1343, 25–27 Aug. 2009.
[12] G. E. Hinton, S. Osindero and Y. W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* **18** (2006), 1527–1554.
[13] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kings-bury, Deep neural networks for acoustic modeling in speech recognition, *Signal Process. Mag.* **29** (2012), 82–97.
[14] Y. Hu and P. Loizou, Subjective comparison and evaluation of speech enhancement algorithms, *Speech Commun*. **49** (2007), 588–601.
[15] Y. Hu and P. C. Loizou, Evaluation of objective quality measures for speech enhancement, *IEEE Trans. Audio Speech Lang. Process*. **16** (2008), 229–238.
[16] M. Jansen, Noise reduction by wavelet thresholding, in: *Ser. Lecture Notes in Statistics*, vol. 161, Springer-Verlag, Berlin, Germany, 2001.
[17] S. Kamath and P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* Orlando, USA, May 2002.
[18] H. Liu, X. Yu, W. Wan and R. Swaminathan, An improved spectral subtraction method, in: *Int. Conf. on Audio, Language and Image Processing (ICALIP)*, Shanghai, pp. 790–793, July 2012.
[19] P. C. Loizou, Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum, *IEEE Trans. Speech Audio Process.* **13** (2005), 857–869.
[20] P. Loizou, *Speech enhancement: theory and practice*, 1st ed., CRC Taylor & Francis, Boca Raton, FL, 2007.
[21] T. Lotter and P. Vary, Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model, *EURASIP J. Appl. Signal Process.* **5** (2005), 1110–1126.
[22] Y. Lu and P. C. Loizou, Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty, *IEEE Trans. Audio Speech Lang. Process.* **19** (2011), 1123–1137.
[23] S. Mallat, A wavelet tour of signal processing, Academic Press, San Diego, CA, 1999.
[24] R. Martin, Speech enhancement based on minimum mean-square error estimation and supergaussian priors, *IEEE Trans. Speech Audio Process.* **13** (2005), 845–856.
[25] iITU-T, Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone net-works and speech codecs, ITU, ITU-T Rec. P. 862, ITU-T, Geneva, Switzerland, 2000.
[26] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiìat, A. Rastrow, R. C. Rose, P. Schwarz and S. Thomas, The subspace gaussian mixture model-a structured model for speech recognition, in: *Computer Speech and Language*, pp. 404–439, Elsevier, Amsterdam, The Netherlands, 2011.
[27] L. R. Rabiner, Applications of voice processing to telecommunications, *Proc. IEEE* **82** (1994), 199–228.
[28] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall, Inc, Upper Saddle River, NJ, USA, 1993.
[29] J. Ramirez, J. M. Gorriz and J. C. Segura, Voice activity detection. Fundamentals and speech recognition system robust-ness, in: *Robust Speech Recognition and Understanding*, M. Grimm, K. Kroschel, eds., ISBN 987-3-90213-08-0, pp. 460, I-Tech, Vienna, Austria, 2007.
[30] A. Rix, J. Beerends, M. Hollier and A. Hekstra, Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs, in: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*. vol. 2, pp. 749–752, 2001.
[31] R. C. Rose, S. C. Yin and Y. Tang, An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition, in: *Proc. ICASSP*, pp. 4508–4511, Prague, Czech Republic, 2011.
[32] G. Y. Thimmaraja and H. S. Jayanna, A spoken query system for the agricultural commodity prices and weather information access in Kannada language, *Int. J. Speech Technol. Springer* **20** (2017), 635–644.
[33] A. Trihandoyo, A. Belloum and K. M. Hou, A real-time speech recognition architecture for a multi-channel interactive voice response system, *Proc. ICASSP* **4** (1995), 2687–2690.
[34] D. Wang and G. Brown, Eds., Computational auditory scene analysis (CASA): principles, algorithms, and applications, Wiley/IEEE Press, Piscataway, NJ, 2006.
[35] P. J. Wolfe and S. J. Godsill, Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement, in: *Proc. 11th IEEE Signal Process. Workshop Statist. Signal Process*., pp. 496–499, Singapore, Aug. 2001.
[36] B.-Y. Xia, Y. Liang and C.-C. Bao, A modified spectral subtraction method for speech enhancement based on masking prop-erty of human auditory system, in: *Int. Conf. on Wireless Communications Signal Processing, WCSP,* pp. 1–5, Nanjing, China, Nov. 2009.