International Journal of Speech Technology (2020) 23:149–167
https://doi.org/10.1007/s10772-020-09671-5

Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling

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

Received: 19 July 2018 / Accepted: 2 January 2020 / Published online: 22 January 2020

Abstract
In this paper, the improvements in the recently implemented Kannada speech recognition system is demonstrated in detail. The Kannada automatic speech recognition (ASR) system consists of ASR models which are created by using Kaldi, IVRS call flow and weather and agricultural commodity prices information databases. The task specific speech data used in the recently developed spoken dialogue system had high level of different background noises. The different types of noises present in collected speech data had an adverse effect on the on line and off line speech recognition performances. Therefore, to improve the speech recognition accuracy in Kannada ASR system, a noise reduction algorithm is developed which is a fusion of spectral subtraction with voice activity detection (SS-VAD) and minimum mean square error spectrum power estimator based on zero crossing (MMSE-SPZC) estimator. The noise elimination algorithm is added in the system before the feature extraction part. An alternative ASR models are created using subspace Gaussian mixture models (SGMM) and deep neural network (DNN) modeling techniques. The experimental results show that, the fusion of noise elimination technique and SGMM/DNN based modeling gives a better relative improvement of 7.68% accuracy compared to the recently developed GMM-HMM based ASR system. The least word error rate (WER) acoustic models could be used in spoken dialogue system. The developed spoken query system is tested from Karnataka farmers under uncontrolled environment.

Keywords Speech · Speech recognition · Interactive voice response system (IVRS) · Automatic speech recognition (ASR) · Spectral subtraction with voice activity detection (SS-VAD) · Minimum mean square error spectrum power estimator based on zero crossing (MMSE-SPZC) · Minimum mean square error spectrum power (MMSE-SP) · Maximum a Posteriori (MAP)

1 Introduction

There are 6.5 crore people disseminated over Karnataka state under different dialect regions and they are daily deals with different commodities Karnataka Raitha Mitra (2008). The agricultural marketing network (AGMARKNET) website is maintaining by agricultural ministry, Government of India provides agricultural commodity price information for Indian languages Agricultural Marketing Information Network (2011). This website is updated frequently and provides minimum modal and maximum price information of particular commodity in different Indian languages. Many farmers in Karnataka state are uneducated and do not computer savvy but almost all farmers uses mobiles for their interaction purpose India Telecom Online (2013). Therefore, it is less cost to combine the mobile network with automatic speech recognition (ASR) system. Integrating the mobile network with ASR models to build spoken query system gives a good result for the statement of problem demonstrated in Kotkar et al. (2008), Rabiner (1997). An end to end spoken dialogue system consists of three main steps. They are, interactive voice response system (IVRS) call flow, ASR models developed from Kaldi speech recognition toolkit and AGMARKNET commodity price information database management system. The IVRS call flow structure is used for task specific speech data collection. The PHP programming language is used to develop the call flow for the speech data collection and speech recognition system. The ASR acoustic

✉ G. Thimmaraja Yadava
thimrajyadav@gmail.com

[1] Department of Electronics and Communication Engineering, School of Engineering and Technology, Jain Deemed to be University, Kanakapura Road, Karnataka, India

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

models are created using Kaldi. In Shahnawazuddin et al. (2013), an Assamese spoken dialogue system is demonstrated to spread the price information of agricultural commodities in Assamese language/dialects. The acoustic ASR models were developed by using speech data which was gathered from the real farmers of Assam. The constrained speech data unseen speaker adaptation method was derived and it was known to give a significant development by 8% over initial evaluation. In Ahmed et al. (2014), an Arabic ASR system is developed using Arabic language resources and data sparseness. The basic modeling techniques such as Gaussian mixture model (GMM) and hidden Markov model (HMM) were used to build an acoustic ASR models for Arabic speech recognition system.

The 36 phonetic symbols and 200 h of speech corpus were used. A Russian ASR system was developed using syntactico-statistical modeling algorithm with big Russian dictionary Karpov et al. (2014). The standard IPA phonemes were used as quality phoneme set to build ASR models. It includes phonetic symbols of 55, consonants symbols of 38 and vowels of 17 in dictionary. The Russian language speech corpus was recorded in clean environment. The 16 kHz sampling rate and 26 h of speech corpus was given for Kaldi system training and decoding. The obtained word error rate (WER) was 26.90%. The improvements in Assamese spoken dialogue system are implemented in Dey et al. (2016). Foreground speech segmentation enhancement algorithm was used for the suppression of different background noises. The noise elimination algorithm was introduced before the Mel frequency cepstral coefficients (MFCC) features extraction part. Recently introduced modeling techniques such as subspace Gaussian mixture model (SGMM) and deep neural networks (DNN) were used for the development of acoustic models. The developed spoken dialogue system was verified by the farmers of Assam under degraded condition and it enables the farmers/users to obtain the on-time price information of agricultural commodities in Assamese language/dialects.

The ASR system using IVRS is one of the stupendous applications of speech processing Rabiner (1994). The amalgamation of IVRS and ASR systems are called spoken query systems which are used to decode the user input Glass (1999), Trihandoyo et al. (1995) 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 prices of agricultural commodities and weather information for Kannada language/dialects Thimmaraja and Jayanna (2017). This work is an ongoing sponsored project by DeitY, Government of India, targeted to develop an user friendly spoken dialogue system by addressing the needs of Karnataka farmers. The developed system gives an option to the user to make his/her own query about what he/she wants over mobile/land-line telephone network. The query which is uttered by user/farmer is recorded, checked the price/weather information in database through ASR models and communicate the on time price/weather information of particular commodity in particular district through saved messages (voice prompts). The earlier spoken query system Thimmaraja and Jayanna (2017) was developed using GMM-HMM modeling techniques. In this work, we demonstrate an enhancements to the recently implemented ASR system. A noise elimination algorithm is proposed which is used to reduce the noise in speech data collected under uncontrolled environment. We have also investigated the two different acoustic modeling approaches reported latterly subjected to spoken query system. The training and testing speech data used in Thimmaraja and Jayanna (2017) for the creation of ASR models was collected from the farmers under real time environment. Therefore, the collected speech data was adulterated by different types of background noises and when the user/farmer makes a query to the system also happens to have high level of background noise. This totally decrease the entire spoken query system performance. To overcome this problem, we have introduced the noise elimination algorithm before 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 process of enhancing the degraded speech data using various noise elimination techniques is called speech processing Loizou (2007). The modified spectral subtraction algorithm was proposed for improvements in speech Bingyin et al. (2009). This algorithm was implemented by using MCRA method Cohen and Berdugo (2002). The conducted experimental results analysis were evaluated and compared with existing methods. In Liu et al. (2012), to suppress the musical noise in corrupted speech data, a modified spectral subtraction algorithm was proposed and it was compared with traditional spectral subtraction algorithm.

Few years back, two advanced acoustic modeling approaches namely SGMM and DNN have been described in Povey et al. (2011), Dahl et al. (2012), Hinton et al. (2012) and Hinton et al. (2006). These two techniques provide better speech recognition performance than GMM-HMM based approach. The GMM in acoustic space is called as SGMM. Therefore, the SGMM is best suitable for the moderate training speech data. Furthermore, DNN consists of more hidden layers in multi layer perception to capture nonlinearities of training set. This gives a good improvement in modeling of variations of acoustics leading to better performance of speech recognition. The process of identifying the human

speech in to its equivalent text format is called speech recognition. The speech recognition output can be used in various applications. Nowadays many artificial intelligence techniques are available to build robust ASR models for the development of end to end ASR systems Derbali et al. (2012). The authors have used the Sphinx toolkit to model the sequential structure and its classification of patterns. The recently developed speech recognition toolkit is the Sphinx-4 which is an added value to CMU repository. It was jointly implemented by various universities and laboratories. The Sphinx-4 is having more advantages compared earlier versions Sphinx systems. They are in terms of flexibility, reliability, modularity and algorithmic aspects. Sphinx-4 supports different languages/dialects and it uses different searching strategies. The entire package of Sphinx-4 is developed by using JAVA programming language. It is very user friendly, portable, flexible and very easy to work with multi threading concepts Lamere et al. (2003).

The implementation and design of natural Arabic continuous speech recognition system was developed in Abushariah et al. (2010) using Sphinx tool. The authors have explored the effectiveness of Sphinx models and developed a robust new continuous Arabic ASR system. The newly developed Arabic speech recognition system is compared with the baseline Arabic ASR. The implemented ASR system was used the Sphinx and HTK tools to build language and acoustic models. The speech signal feature vectors are extracted from MFCC technique. The system was used five state HMM and 16 GMMs. The number of senons used in this work are 500 and the developed model used 7 h of transcribed and validated speech data. One hour of speech data was used for decoding purpose. The achieved accuracy of speech recognition is 92.67%.

The ASR system was developed and designed in Al-Qatab and Ainon (2010) using HTK tool. The system was developed for both isolated and continuous speech recognition purpose. The lexicon and phoneme set was created first for Arabic language. The MFCC technique was used for the speech signal features extraction. The speech database was collected from the native Arabic speakers. The achieved speech recognition accuracy was 97.99%. The three important components play an important role in the development of ASR system. They are, lexicon or dictionary, language model and acoustic model. The authors have developed a robust ASR system without using lexicon for English language Harwath and Glass (2014). The impact of subspace based modeling techniques is investigated in Rose et al. (2011). The SGMM is the new level of modeling technique to build acoustic models for the development of robust ASR systems. The SGMM technique was used to develop continuous speech recognition system and achieved WER was 18%. An ASR system was developed for the Odia language in Karan et al. (2015). The ASR models were developed by using Kaldi. The speech database was collected from the farmers of Odisha in real time environment. The ASR models were developed for the district names of Odisha state. The ASR models were developed by using monophone and triphone training and decoding. The asterisk server was used to develop an ASR system. The enhancements in IITG spoken query system was described in Abhishek et al. (2017). The models were built by using Kaldi. An end to end spoken dialogue system consist of IVRS flow, IMD and AGMARKNET databases and ASR models. The earlier IITG spoken dialogue system leads to a maximum WER due different noises were present in the collected speech data. In order to overcome from the problem less accuracy, the authors have developed a robust noise elimination algorithm and introduced it before the MFCC features extraction part. The earlier developed system was modeled by using GMM-HMM based technique. To improve the accuracy, the authors have built the ASR models using SGMM-DNN base modeling technique. The comparison of on line and off line speech recognition accuracies is also done in their work. An added version of Sphinx called Sphinx-4 framework was developed in Walker et al. (2004). The developed Sphinx-4 version is very robust in the development of language and acoustic models. The Sphinx-4 is modular, extend, portable and flexible. The Sphinx-4 and its additional packages are freely available on the Internet. The continuous ASR system for large vocabulary based on DNN was presented in Popovic (2015). The ASR models were built by using Kaldi. The DNNs are mainly implemented based on the principle of restricted Boltzmann machines. The GMMs were used to represent the HMM state's emission densities. 90 h of continuous Serbian speech data was used for training the system. The performances of GMM-HMM based models and DNN models were compared and the best model could be used in continuous ASR system. The improvements in Arabic ASR system was presented in Nahar and Squeir (2016). The authors have introduced a novel hybrid approach to increase the accuracy of speech recognition. The hybrid technique was a combination of learning vector quantization and HMM. This algorithm was mainly intended to recognize the phonemes in the continuous speech vocabulary. The TV news speech corpus was taken for system training and testing. Therefore the achieved speech recognition accuracies are 98.49% and 90% for independent and dependent training respectively. A new modeling technique was developed Ansari and Seyyedsalehi (2016)for the development of robust ASR models. The modular deep neural network was introduced for the recognition of speech. The pre-training of network is mainly depends on its structure. The two important speaker adaptation techniques were also implemented in this work. For the system training and testing, two speech databases were used and achieved WERs were 7.3% and 10.6% for MDNN

and HMM respectively. The ASR models were built by using recurrent neural networks for Russian language was presented in detail Kipyatkova and Karpov (2017). These ANNs were used for the development of robust continuous ASR system for Russian language. The models of neural network are constructed based on the principle of number of elements in the hidden layer. An unsupervised learning models were introduced in Sailor and Patil (2016) which are based on RBMs. The authors have experimented both MFCC features and filterbank features on large vocabulary of continuous speech data. An AURORA-4 speech dataset was used for the experimental conduction. The proposed filterbank provides better performance compared to traditional MFCC feature extraction technique. The traditional methods for speech enhancement based on different frames or segments needs much knowledge about different noises. A new algorithm was introduced for the elimination of different types of noises in the corrupted speech data Ming and Crookes (2017). The authors have used zero mean normalized correlation coefficient as a measure of comparison. This algorithm overcomes the problem of knowing the knowledge about different noises present in the speech data. The proposed method outperforms the conventional traditional speech enhancement methods. The performances of conventional and proposed methods were done by considering some objective measures as well as ASR. A speech enhancement algorithm called missing feature theory was developed in Van Segbroeck and Van Hamme (2011) to improve the accuracy of ASR system. The missing feature theory algorithm was applied on log spectral domain and static and dynamic features. To compute the channel, a maximum likelihood computation technique was integrated with missing feature theory. The Aurora-4 speech database was used for experimental conduction. For the structured classification work, a discriminative models were used for speech recognition Zhang et al. (2010). The authors have developed a set structured log linear models to develop a robust ASR system. The main advantage of log linear models is its features. The proposed method was the combination of kernels, efficient lattice margin training, discriminative models and noise compensation technique. An Aurora-2 speech database was used for the experimental conduction.

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 and created entire dictionary and phoneme set for 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 dialogue system.
- Testing the newly developed spoken dialogue system from farmers of Karnataka under uncontrolled environment.

The rest of the work is summarized as follows: Sect. 2 describes the collection speech database and its preparation. The background noise elimination by fusing SS-VAD and MMSE-SPZC estimator is described in Sect. 3. The creation of ASR models using Kaldi is described in Sect. 4. The effectiveness of SGMM and DNN is described in Sect. 5. The experimental results and analysis are discussed in Sect. 6. The Sect. 7 gives the conclusions.

2 Speech database collection from farmers

The basic building block diagram of different steps involved in the development of improved Kannada spoken dialogue system is given in Fig. 1.

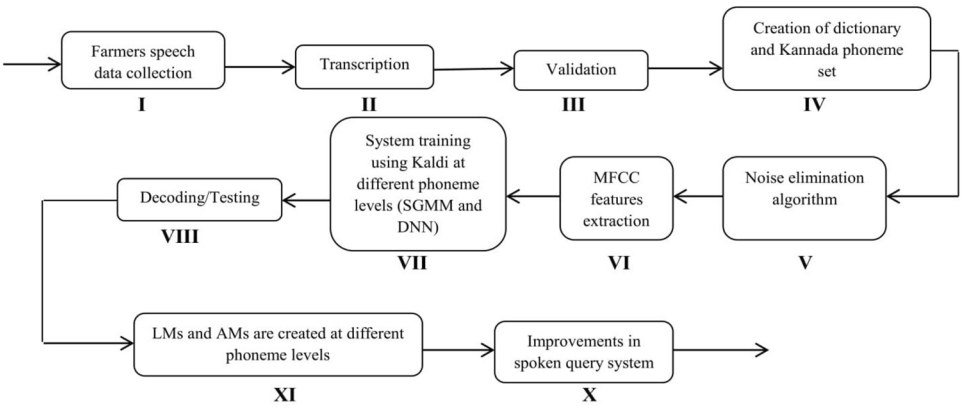

Fig. 1 Block diagram of different steps involved in the development of Kannada spoken query sytem

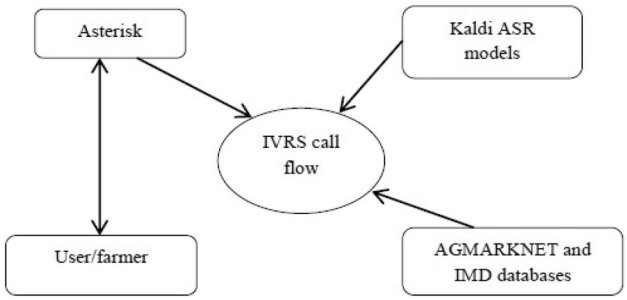

Fig. 2 An integration of ASR system and Asterisk

An Asterisk server is used in this work which acts as an interface between the user/farmer and the IVRS call flow is shown in Fig. 2.

To increase the speech database, another 500 farmers speech data is collected in addition to the earlier speech database. The training and testing speech data set consists of 70757 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 AGMARKNET list under Karnataka section. The performance estimation of entire ASR system is done by overall speech data.

3 Combination of SS-VAD and MMSE-SPZC estimator for speech enhancement

A noise reduction technique is proposed for speech enhancement which is an amalgamation of SS-VAD and MMSE-SPZC estimator. Consider an original speech signal s(n) which is degraded by background noise d(n). Therefore, the degraded speech c(n) can be represented as follows:

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

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 interval which can be given as

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

the STFT of c(n) can be written as

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

the polar form representation of the above equation can be written as follows:

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

where $\{C_k, S_k, D_k\}$ represents the magnitudes and $\{\theta_c(k), \theta_s(k), \theta_d(k)\}$ represents the phase. The power spectrum of degraded speech signal can be written as follows:

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

The equation (6) can be written as follows:

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

The *pdf* of S_k^2 and D_k^2 are given 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 posteriori probabilities of speech signal magnitude squared spectrum is evaluated by using Bayes theorem as shown below.

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

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

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

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

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 method with VAD

The block diagram of spectral subtraction algorithm is given in Fig. 3. The spectral subtraction algorithm is mainly incorporated with VAD which is used to find the voiced regions in

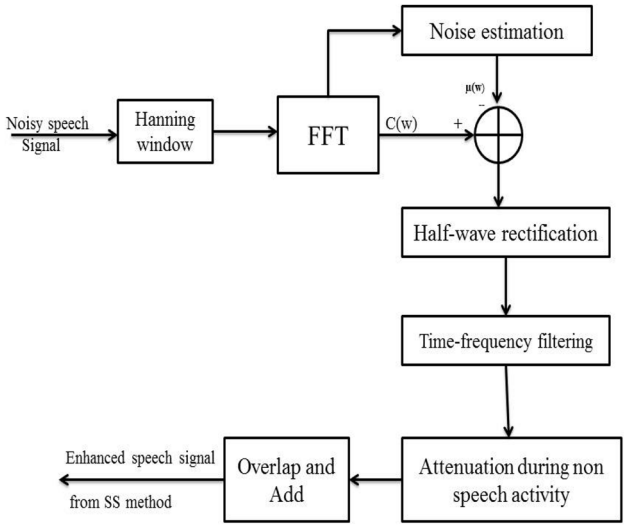

Fig. 3 Schematic representation of SS-VAD method

the speech signal. Consider s(n), d(n) and c(n) are the original, additive noise and corrupted speech signal respectively. The steps shown in the Fig. 3 are followed neatly to enhance the corrupted speech data. The linear prediction error can also be called as L which is mainly incorporated with Energy of the speech signal E and zero crossing rate (Z). Therefore, the term Y can be represented as follows:

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

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

The output of spectral subtraction with VAD algorithm can be represented as follows:

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

The negative values of output speech spectrum are set to zero using half wave rectification if they have negative values. In order to attenuate the signal further during non speech activity, the residual noise reduction process is used. It improves the enhanced speech signal quality.

3.2 MSS estimators

Te following three types of magnitude squared spectrum estimators (MSSE) are studied, implemented and their performances are compared.

3.2.1 MMSE-SP estimator

In Wolfe and Godsill (2001), authors have proposed a technique called MMSE-SP estimator. The clean speech signal can be written as follows:

$$\hat{S_K^2} = E\left\{S_k^2 | C(w_k)\right\} \tag{18}$$

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

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

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 function $f_{(S_k)}(S_k|Y(w_k))$ can be written 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

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

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

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

3.2.2 MMSE-SPZC estimator

An another important magnitude squared spectrum estimator is MMSE-SPZC. By using the work which was presented in Jounghoon and Hanseok (2003) and Cole et al. (2008), the MMSE-SPZC estimator is derived Lu and Loizou (2011).

$$\hat{S_k^2} = E\left\{S_K^2 | C_k^2\right\} \tag{26}$$

$$\hat{S_K^2} = \int_0^{C_k^2} S_k^2 f_{S_k^2}\left(S_k^2 | C_k^2\right) dS_k^2 \tag{27}$$

$$\hat{S_K^2} = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right) C_k^2, & if\ \sigma_s^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} C_k^2, & if\ \sigma_s^2(k) = \sigma_d^2(k) \end{cases} \tag{28}$$

where v_k can be written as

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

The gain equation of the above estimator can be represented mathematically as follows:

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

3.2.3 MAP estimator

The MAP estimator can be written as shown below:

$$\hat{S_k^2} = argmax f_{S_K^2}\left(S_k^2 | C_k^2\right) \tag{31}$$

maximization with respect to S_K^2.

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

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

The MAP estimator gain function can be represented as follows:

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

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

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

3.3 Measures of performance and analysis

The standard measures are used to evaluate the performances of proposed and existing speech enhancement methods. They are composite measures and perceptual evaluation of speech quality (PESQ).

The scales of ratings of composite measures is shown in Table 1. The three important composite measures Hu and Loizou (2007) are,

- Speech signal distortion (s).
- Background noise distortion (b).
- Overall speech signal quality (o).

3.4 Performance evaluation of existing methods

The TIMIT and Kannada speech databases are used for the experiments conduction. The speech sentences are corrupted by musical, car, babble and street noises. The evaluation of performances of existing and proposed methods are presented in this section.

3.4.1 SS-VAD method performance evaluation

The Tables 2 and 3 shows the performance evaluation of SS-VAD technique in terms of PESQ for both databases. There is a less suppression of musical noise for both databases using SS-VAD technique. The performance evaluation of the same method using composite measures for both databases is shown in Tables 4 and 5. It can be inferred from the tables that there is a less improvement in the speech databases which are degraded by musical noise.

3.4.2 Performance evaluation of MSS estimators

The Tables 6 and 7 shows the performance measurement of MSS estimators using PESQ for both databases. The speech quality is very less for the speech databases which are degraded by babble noise compared to other types of noises as shown in Tables 8 and 9. Among three MSS estimators, the MMSE-SPZC estimator has given better performance for the degraded speech data shown in Tables 6, 7, 8 and 9. Therefore, from the experimental analysis, it can be inferred that, an SS-VAD method gives poor results for the speech database which is degraded by musical noise and MMSE-SPZC technique gives poor speech quality for the database which is degraded by babble noise. The speech databases which are degraded by both musical and babble noises can be easily eliminated by combining an SS-VAD and MMSE-SPZC estimator.

3.5 Combined SS-VAD and MMSE-SPZC estimator for speech enhancement

The flowchart of the proposed method is shown in Fig. 4. The output of SS-VAD is given as input to an MMSE-SPZC estimator because of less improvement in musical noise.

The output of SS-VAD can be represented as follows:

$$|X_i(w)| = |C_i(w)| - |\mu_i(w)| \tag{36}$$

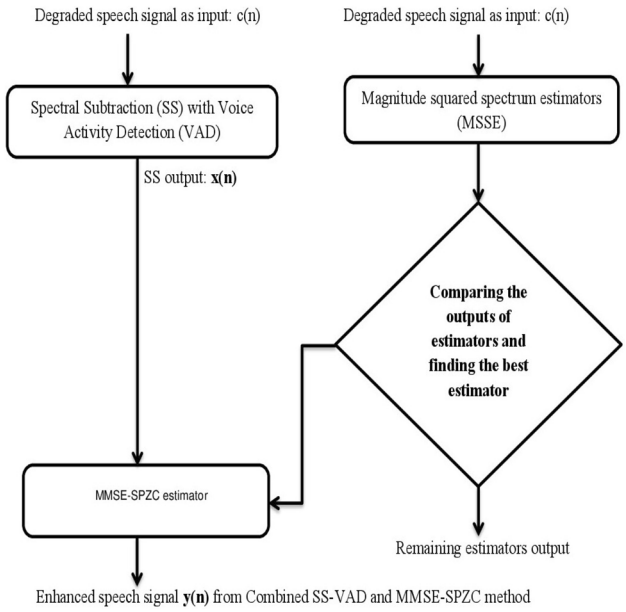

Fig. 4 Flow chart of proposed combined SS-VAD and MMSE-SPZC estimator for speech enhancement

The MMSE-SPZC estimator receives the output of SS-VAD and it can be derived by considering its *pdf* is shown in below equation.

$$\hat{X}_k^2 = E\left\{X_K^2 | Y_k^2\right\} \tag{37}$$

$$\hat{X}_K^2 = \int_0^{X_k^2} X_k^2 f_{X_k^2}(X_k^2 | Y_k^2) dX_k^2 \tag{38}$$

$$\hat{X}_K^2 = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right) Y_k^2 & if\ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} Y_k^2 & if\ \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{39}$$

The gain function of proposed estimator can be represented as follows:

$$G_{MMSE}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{v_k} - \frac{1}{e_k^v - 1}\right)^{\frac{1}{2}} & if\ \ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \left(\frac{1}{2}\right)^{\frac{1}{2}} & if\ \ \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{40}$$

The Tables 10 and 11 shows the experimental results of proposed method using PESQ for both databases. The speech

Table 1 The scales of ratings of composite measures

Ratings	Scale of speech signal (s)	Scale of background noise (b)	Overall scale (o)
1	Much corrupted	Very intrusive	Much poor
2	Fairly corrupted	Fairly intrusive and conspicuous	Poor
3	Medium	Not intrusive and can be noticeable	Medium fair
4	Fairly common with some corruption	Little noticeable	Good
5	No degradation	Can not noticeable	Excellent

Table 2 Evaluation of performance of SS-VAD technique in terms of PESQ for TIMIT database

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

Table 3 Evaluation of performance of SS-VAD technique in terms of PESQ for Kannada database

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

Table 4 Evaluation of performance of SS-VAD method in terms of composite measure for TIMIT database

Method	Composite measure	Musical	Car	Babble	Street
SS-VAD	(s)	1.8017	3.6399	3.5213	2.7860
	(b)	2.6670	2.2760	2.1245	1.9125
	(o)	**3.1639**	3.7759	3.4245	3.3182

Table 5 Evaluation of performance of SS-VAD method in terms of composite measure for Kannada database

Method	Composite measure	Musical	Car	Babble	Street
SS-VAD	(s)	1.9017	3.1199	3.4313	2.2460
	(b)	2.2370	2.2178	2.1245	1.9355
	(o)	**2.9039**	3.6659	3.1244	3.2112

Table 6 Evaluation of performance of MSS estimators in terms of PESQ for TIMIT database

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

Table 7 Evaluation of performance of MSS estimators in terms of PESQ for Kannada database

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

Table 8 Evaluation of performance of MSS estimators in terms of composite measure for TIMIT database

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

Table 9 Evaluation of performance of MSS estimators in terms of composite measure for Kannada database

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

Table 10 Evaluation of performance 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 Evaluation of performance 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 Evaluation of performance of combined SS-VAD and MMSE-SPZC method in terms of composite measure for TIMIT database

Method	Composite measure	Musical	Car	Babble	Street
Proposed method	(s)	3.1204	3.6319	3.5689	2.6052
	(b)	2.6920	2.4780	2.8569	1.9098
	Overall (o)	4.4409	4.3002	4.2956	4.3141

Table 13 Evaluation of performance of combined SS-VAD and MMSE-SPZC method in terms of composite measure for Kannada database

Method	Composite measure	Musical	Car	Babble	Street
Proposed method	(s)	3.1204	3.6319	3.5689	2.6052
	(b)	2.6920	2.4780	2.8569	1.9098
	(o)	4.5111	4.2911	4.3123	4.4112

quality and its intelligibility is improved after the speech enhancement using proposed method in terms of composite measure is shown in Tables 12 and 13. The experimental results show that the SS-VAD algorithm yielded better results for the speech data corrupted by background noise, vocal noise, car noise, street noise and babble noise but not for musical noise. The MMSE-SPZC estimator has given better results for the speech data corrupted by musical noise, background noise, and street noise but not for babble noise. This is due to the collected speech data was much degraded by musical and babble noises. Therefore, in order to improve the quality of speech data which was degraded by babble, musical and other types of noises, we have combined both algorithms. From the tables, it can be inferred that, the proposed method significantly reduced the babble, musical and other different types of noises in both speech databases compared to individual methods. Hence the proposed noise elimination algorithm could be used for speech enhancement in Kannada spoken query system to improve the speech recognition accuracy under uncontrolled environment.

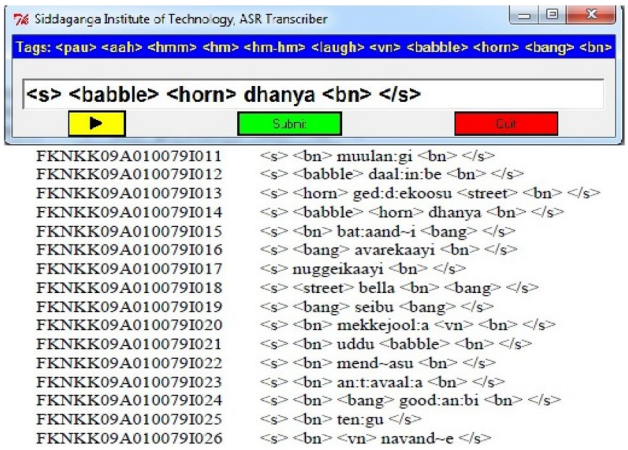

Fig. 5 Transcription of speech data

4 Creation of ASR models using Kaldi for noisy and enhanced speech data

The development of ASR models includes several steps. They are as follows:

- Transcription and validation of enhanced speech data.
- Development of lexicon and Kannada phoneme set.
- Extraction of MFCC features.

4.1 Transcription and validation

Transcription is a method of converting speech file content in to its equivalent word format which can also called as word level transcription. The schematic representation of

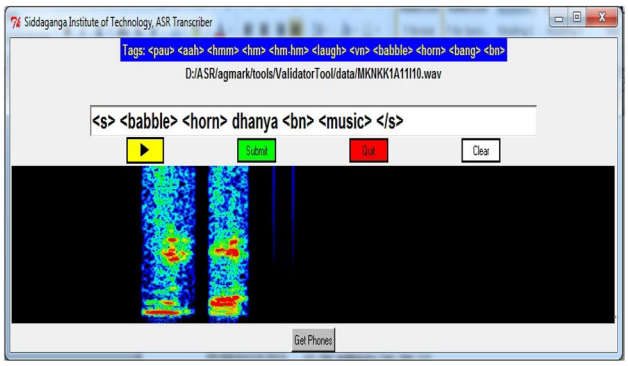

Fig. 6 Validation tool to validate the transcribed speech data

various speech wave files and those equivalent transcription is shown in Fig. 5.

It was observed in the figure that, the tags < s > and < /s > indicates that 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 particular speech data is done wrongly then it will be validated using validation tool is shown in Fig. 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 data, the validator has listen 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 >** is shown in Fig. 6.

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	ಓ	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	ಓ	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

4.2 Kannada phoneme set and corresponding dictionary creation

The Karnataka is one of the states in India. There are 6 Crore people lives in Karnataka state who are fluently speak Kannada language. Kannada language has 46 phonetic symbols and it is one of the most usable Dravidian languages. The description of Kannada phoneme set, Indian language speech sound label 12 (ILSL12) set used for Kannada phonemes and its corresponding dictionary is shown in Tables 14, 15 and 16 respectively.

4.3 MFCC features extraction

Once the noise elimination algorithm is applied on 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 MFCC features extraction is shown in Fig. 7. The Kannada speech data used for the experimental conduction is huge and most of the speech information is present in the first 13 coefficients. Therefore, to minimize the complexity of algorithm, 13 dimensional MFCC coefficients have been used in this work.

The parameters used for MFCC features extraction are as follows:

- Window used: Hamming window.
- Window length: 20 ms.
- Pre-emphasis factor: 0.97.
- MFCC coefficients: 13 dimensional.
- Filter bank used: 21 Channel filter bank.

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
daavand~agere	d aa v nx a g e r e
daavand~agere	d aa v a nx g e r e
tumakuuru	t u m a k uu r u
tumakuuru	t u m a k uu r
tumakuuru	t u m k uu r u
tumakuuru	t u m u 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
ben:gal:uuru	b e ng lx uu r u
ben:gal:uuru	b e ng lx uu r
chaamaraajanagara	c aa m a r aa j a n a g a r a
chaamaraajanagara	c aa m r aa j a n a g a r a
chaamaraajanagara	c aa m r aa j n a g a r a
shivamogga	sh i v a m o gg a
shivamogga	sh i v m o gg a
shivamogga	sh i m o gg a
haaveiri	h aa v ei r i
gulbarga	g u l b a r g a
gulbarga	g u l b a r g
dhaarawaad:a	d aa r a v aa dx a
dhaarawaad:a	d aa r a v aa dx
raamanagara	r aa m a n a g a r a
raamanagara	r aa m n a g a r a
raamanagara	r aa m a n a g a r
raamanagara	r aa m n a g r a
koolaara	k oo l aa r a
koppal:a	k o pp a lx a
koppal:a	k o pp lx a
raayacuuru	r aa y a c uu r u
raayacuuru	r aa y c uu r u
raayacuuru	r aa y c uu r
man:gal:uuru	m a ng g a lx uu r u
man:gal:uuru	m a ng lx uu r u
man:gal:uuru	m a ng lx uu r
man:d:ya	m a n dx y a
cikkamagal:uuru	c i kk a m a g a lx uu r u
cikkamagal:uuru	c i kk a m a g lx uu r u
cikkamagal:uuru	c i kk m a g a lx uu r u
cikkamagal: uuru	c i kk m a g lx uu r
ud:upi	u dx u p i
ud:upi	u dx p i

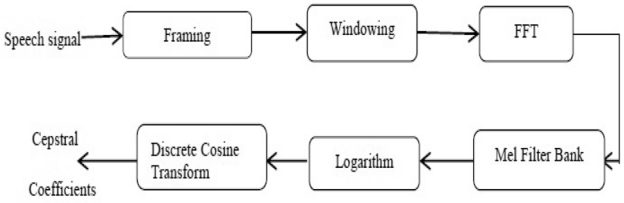

Fig. 7 Block digram of MFCC features extraction

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 is introduced to the speech recognition domain is called SGMM Povey et al. (2011). 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 distributes a usual structure between the states in SGMM modeling technique. The SGMM consist of 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 SGMM can be described by the below equations is as follows:

$$p(y|i) = \sum_{k=1}^{L} w_{ik} N(y;\mu_{ik}, \Sigma_k) \tag{41}$$

$$\mu_{ik} = M_k I_i \tag{42}$$

$$w_{ik} = \frac{expw_k^T I_i}{\sum_{k'=1}^{L} expw_k'^T I_i} \tag{43}$$

where $y \in R^D$ is a feature vector and $i \in \{1....I\}$ is the context-dependent state of speech signal. The speech state j's model is a GMM with L Gaussians (L is in 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 $\mu_{ik} w_{ik}$ parameters are done by using I_i together with M_k, w_k and Σ_k. The detailed description of parameterization of SGMM and its impact is given in Rose et al. (2011). The ASR models are developed using this modeling technique for Kannada speech database and the least word error rate (WER) models could be used in spoken query system.

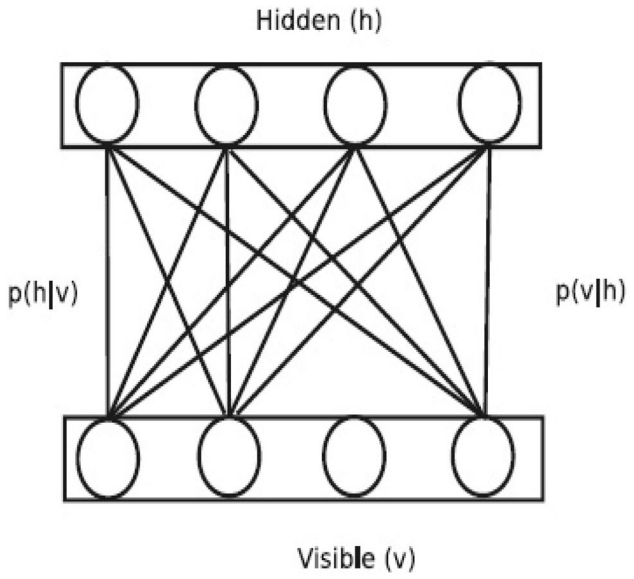

Fig. 8 Block digram of restricted Boltzmann machine

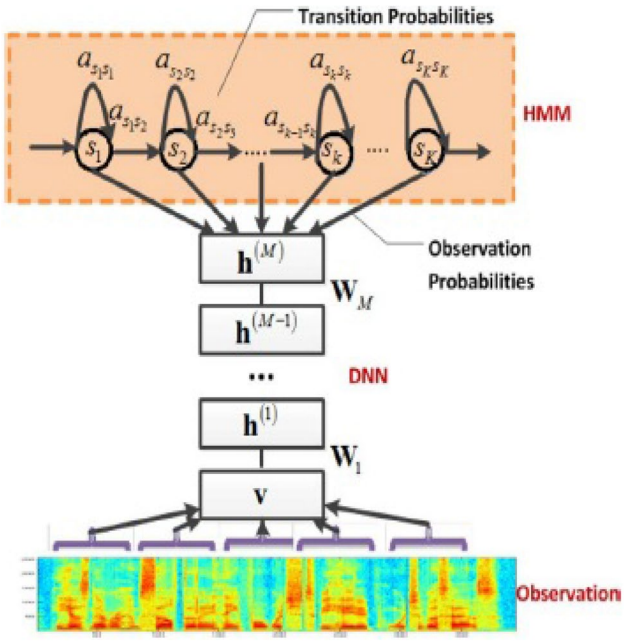

Fig. 9 Block digram of hybrid DNN and HMM

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 GMM-HMM based acoustic modeling approach are discussed in Hinton et al. (2012). 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 maximum number of hidden layers with back propagation algorithm for large amount of speech data. An ANN with single hidden layer failed to give good improvements over the GMM-HMM based acoustic modeling technique. Both the aforementioned limitations have been overcome with the developments in the past few years. Various approaches are available now to train the different neural nets with maximum number of hidden layers.

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

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

$$p(v,h;W) = \frac{1}{Z}e^{-E(v,h;W)} \tag{44}$$

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

$$Z = \sum_{v',h'} e^{-E(v',h';W)} \tag{45}$$

where v' and h' are the extra variables used for summation over the ranges of v and h. The unsupervised technique is described in detail in Hinton et al. (2006) 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 context dependent DNN-HMM hybrid architecture is shown in Fig. 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 DNN is

Table 17 The speech files used for system training and testing

Kannada speech database	Number of train files	Number of test files
Previous work districts utterances	9123	3008
Previous work mandi utterances	15589	5563
Previous work commodities utterances	19550	7520
Previous work overall speech utterances	40,235	10123
Present work overall noisy speech data	68523	2180
Present work overall enhanced speech data	67234	2180

Table 18 The WERs for overall noisy speech data

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 (DNN: Iteration 1)	11.88	11.89	11.89	11.86	11.88	11.82
tri4_nnet_t2a (DNN: Iteration 2)	11.62	11.25	11.88	11.24	11.11	11.25
tri4_nnet_t2a (DNN: Iteration 3)	11.25	11.01	11.05	11.22	11.23	11.56
tri4_nnet_t2a (DNN: Iteration 4)	11.01	10.89	10.81	10.86	10.88	10.80

used with HMM which models the sequential property of the speech.

6 Experimental results and analysis

The number of speech files used for previous and present work for system training and testing are shown in Table 17.

The ASR models are created at different phoneme levels are as follows:

- Monophone, Triphone1 (delta, delta-delta training and decoding), Triphone2 (linear discriminant analysis (LDA)+maximum likelihood linear transform (MLLT)) and 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 ASR models. The recently introduced two modeling techniques such as SGMM and DNN are used to build ASR models. Table 18 shows the description of WERs at different phoneme levels for overall noisy speech data. The WERs of 12.78% and 10.80% are achieved

Table 19 TThe WERs for overall enhanced speech data

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 (DNN: Iteration 1)	10.67	10.69	10.70	10.70	10.71	10.73
tri4_nnet_t2a (DNN: Iteration 2)	10.62	10.59	10.60	10.58	10.71	10.66
tri4_nnet_t2a (DNN: Iteration 3)	10.24	10.44	10.21	10.44	10.21	10.33
tri4_nnet_t2a (DNN: Iteration 4)	10.01	10.02	9.60	10.11	10.08	9.89

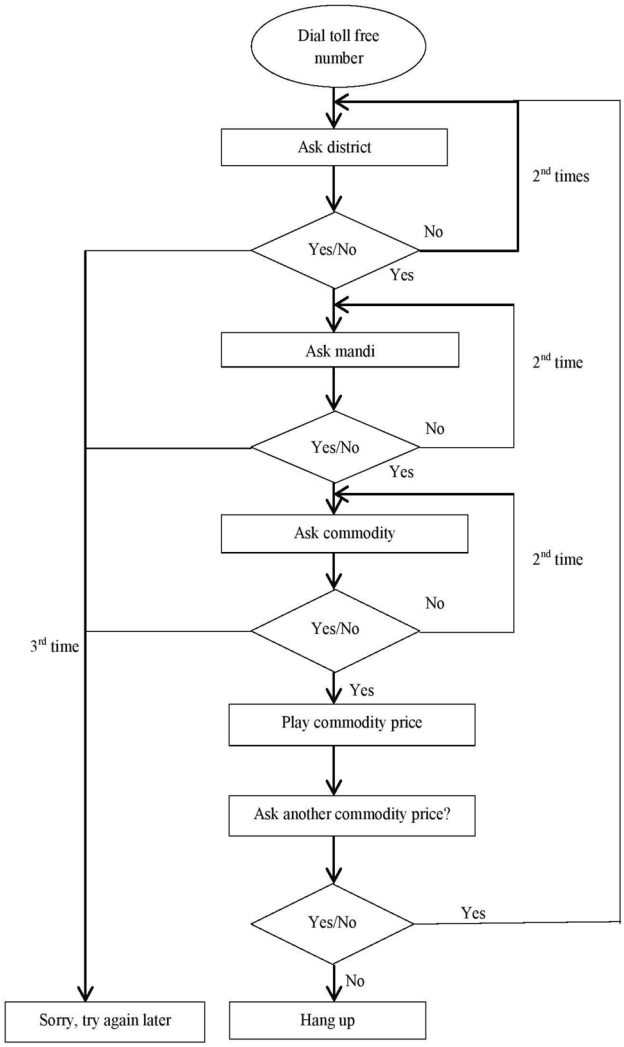

Fig. 10 Call flow structure of commodity price information spoken query system (reproduced from Thimmaraja and Jayanna (2017) for correct flow of this work)

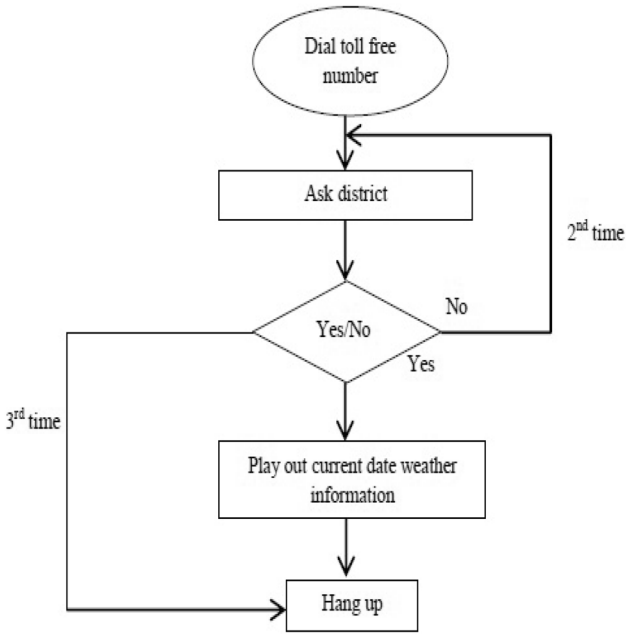

Fig. 11 Call flow structure of weather information spoken query system (reproduced from Thimmaraja and Jayanna (2017) for completeness

for SGMM and hybrid DNN training and decoding respectively for overall noisy speech data.

The WERs of 11.77% and 9.60% is obtained for overall enhanced speech data using SGMM training and decoding and hybrid DNN training and decoding respectively shown in Table 19.

6.1 Call flow structure of spoken dialogue system

The ASR models were developed in the earlier work Thimmaraja and Jayanna (2017) at monophone, triphone1 and triphone 2 levels with 600 senons and 4, 8 and 16 GMMs. The achieved least WERs were 10.05%, 11.90%, 18.40% and 17.28% for districts, mandis, commodities and overall speech data respectively. This leads to the less recognition of commodities and mandis. To overcome this problem, another 500 farmers speech data is collected, increased the training data set and applied noise elimination algorithm on both training and testing data set to improve the accuracy of ASR models in this work. In the earlier work, a separate spoken query systems were developed Thimmaraja and Jayanna (2017). In this work, the two call flow structures have been integrated together and made it as single call flow to overcome the complexity of dialing multiple call flows. Therefore, the user/farmer can access both information in single call flow by dialing toll free number. The earlier call flow structure of spoken query systems are shown in Figs. 10 and 11 respectively.

The schematic representation of new integrated call flow structure is shown in Fig. 12.

The ASR models are developed using monophone, tri1, tri2, tri3 (LDA+MLLT_SAT), SGMM and DNN for overall noisy and enhanced speech data in this work. From the tables, it can be observed that there is better improvement in speech recognition accuracy for enhanced speech data. Approximately, 1.2% of speech recognition accuracy is improved for the speech data after noise reduction. The ASR models are developed for overall speech data to reduce the complexity in the call flow structure of spoken dialogue system. The developed least WER models (overall enhanced speech data models) could be used in spoken dialogue system to improve its on line recognition performance. The Kannada spoken query system need to recognize 250

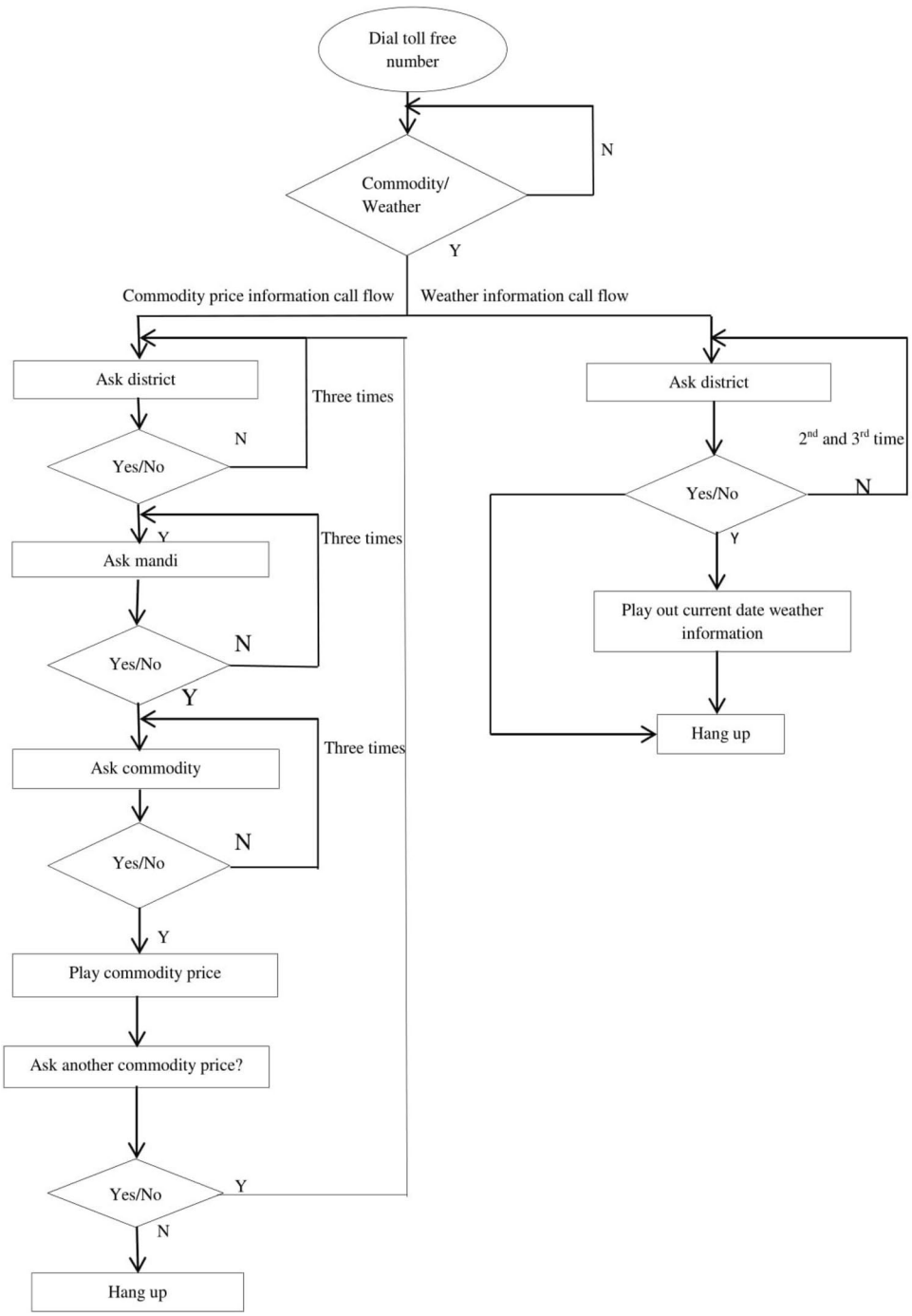

Fig. 12 Call flow structure of spoken query system to access both agricultural commodity prices and weather information

commodities includes all its varieties. The 28 districts, 95 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 mandi name. If the mandi name is also recognized then it will asks the farmer to prompt the commodity name. If the commodity name is recognized then it will play out the current price information of asked commodity from the price information database. Similarly, to get weather information the farmer needs 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 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

Table 20 Performance evaluation of online 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	243	17	10	270	90.00
Mandis	300	242	20	12	270	90.00
Commodities	300	243	17	9	269	89.66

are not recognized properly then the system says sorry! Try once again!.

The following error recovery mechanisms have been incorporated in the spoken query system:

- The voice activity detection is used to detect the poor responses given from the users/speakers. In these iterations, the speakers/farmers are asked to repeat the query or responses more loudly.
- The speaker/farmers are prompted to say yes or no after each and every speech utterance recognition.
- To maximize the level of confidence in the recognized output, three parallel decoders are used to recognize the user response and their outputs are surveyed to generate the final response.

6.2 Testing of developed spoken query system from farmers in the field

The spoken dialogue system is again developed for new ASR models. To check the on line speech recognition accuracy of newly developed spoken dialogue system, the 300 farmers are asked to test the system under uncontrolled environment. The Table 20 shows the performance evaluation of 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 same as shown in Tables 19 and 20.

7 Conclusions

The improvements in recently developed Kannada ASR system is demonstrated in this work. A robust noise elimination technique is proposed for canceling the different types of noises in collected task specific speech data. The proposed method has given a better performance for Kannada and TIMIT speech databases compared to the individual methods for the same databases. An alternative ASR models (SGMM and DNN modeling techniques) were developed by using Kaldi toolkit. The SGMM-DNN ASR models are found to be outperformed compared to the conventional GMM-HMM based acoustic models. There is a better improvement in speech recognition accuracy of 7.68% in DNN-SGMM based acoustic models compared to earlier GMM-HMM based models. The achieved WERs are 12.78% and 10.80% for noisy speech data using SGMM and hybrid DNN based modeling techniques. The WERs of 11.77% and 9.60% are achieved for enhanced speech data using the same modeling techniques. Therefore, it can be inferred that there is a better improvement in accuracy of 1.2% for enhanced speech data compared to noisy speech data. The least WER models (SGMM and DNN based models) could be used in 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 prices/weather information in a single call flow. The Kannada ASR system is tested from the real farmers of Karnataka state under real time environment is also presented in this work.

Acknowledgements This study was supported by Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology, Government of India.

References

Abhishek, D., Shahnawazuddin, S., Deepak, K. T., Siddika, I., Prasanna, S. R. M., & Sinha, R. (2017). Improvements in IITG Assamese spoken query system: Background noise suppression and alternate acoustic modeling. *Journal of Signal Processing Systems*, *88*(01), 91–102.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010). Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. In Computer and Communication Engineering (ICCCE), 2010 International Conference on, Kuala Lumpur, pp. 1–6.

Agricultural Marketing Information Network—AGMARKNET. (2011). http://agmarknet.nic.in.

Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., Glass, J. (Dec 2014). A complete KALDI recipe for building arabic speech recognition systems. Spoken Language Technology Workshop (SLT), IEEE, South Lake Tahoe, NV, pp. 525–529.

Al-Qatab, B. A. Q., & Ainon, R. N. (2010). Arabic speech recognition using Hidden Markov Model Toolkit(HTK). In 2010 International Symposium on Information Technology, Kuala Lumpur, pp. 557–556.

Ansari, Z., & Seyyedsalehi, S. A. (2016). Toward growing modular deep neural networks for continuous speech recognition.

Neural Computing and Applications, *28*, 1177–1196. https://doi.org/10.1007/s00521-016-2438-x.

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

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

Dahl, G., Yu, D., Deng, L., & Acero, A, (2012). Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. In IEEE Transactions on Audio Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award), pp. 30–42.

David, H., & James, G. (2014) Speech recognition without a lexicon—bridging the gap between graphemic and phonetic systems. INTERSPEECH, Singapore, pp. 14–18.

Derbali, M., Mu'Tasem, J., & Taib, M. (2012). A review of speech recognition with Sphinx engine in language detection. *Journal of Theoretical and Applied Information Technology*, *40*(2), 147–155.

Dey, A., Shahnawazuddin, S., Deepak, K. T., Imani, S., Prasanna, S. R. M., & Sinha, R. (2016). Enhancements in Assamese spoken query system: Enabling background noise suppression and flexible queries. In 2016 Twenty Second National Conference on Communication (NCC), pp. 1–6.

Glass, J. R. (1999). Challenges for spoken dialogue systems. In Proceedings of IEEE ASRU workshop.

Goel, S., & Bhattacharya, M. (July 2010). Speech based dialog query system over asterisk pbx server. In 2nd International Conference on Signal Processing Signal Processing Systems (ICSPS), Dalian.

Hinton, G. E., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kings-bury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computer*, *18*, 1527–1554.

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

Huanhuan, L., Xiaoqing, Y., Wanggen, W., & Ram, S. (July 2012) An improved spectral subtraction method. International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, pp. 790–793.

India Telecom Online—ITO. (2013). www.indiatelecomonline.com.

Jounghoon, B., & Hanseok, K. (2003). A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. IEEE International Conference on Multimedia and Expo, vol. 3, pp. I-648–I-651.

Karan, B., Sahoo, J., & Sahu, P. K. (2015). Automatic speech recognition based odia system. International Conference on Microwave, Optical and Communcation Engineering, December 18–20, 2015, IIT Bhubaneswar, India.

Karnataka Raitha Mitra. (2008). raitamitra.kar.nic.in/statistics.html.

Karpov, A., Markov, K., Kipyatkova, I., Vazhinina, D., & Ronzhin, A. (2014). Large vocabulary Russian speech recognition using syntactico-statistical language modeling. *Speech Communications*, *56*(0167–6393), 213–228.

Kipyatkova, I. S., & Karpov, A. A. (2017). A study of neural network russian language models for automatic continuous speech recognition systems. *Automation and Remote Control*, *78*(5), 858–867.

Kotkar, P., Thies, W., & Amarsinghe, S. (April 2008). An audio wiki for publishing user-generated content in the developing world. In HCI for Community and International Development, Florence, Italy.

Lamere, P., Kwok, P., Evandro, B. G., Singh, R., Walker, W., Wolf, P. (2003). The CMU Sphinx-4 speech recognition system. In IEEE International Conference on Acoustics, Speech and Signal Processing.

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

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

Ming, J., & Crookes, D. (2017). Speech enhancement based on full-sentence correlation and clean speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 531–543.

Nahar, K. M. O., & Squeir, M. A. (2016). Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. *International Journal of Speech Technology*, *19*, 495–508.

Popovic, B., Ostrogonac, S., Pakoci, E., Jakovljevic, N., Delic, V. (2015). Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. Berlin: Springer, https://doi.org/10.1007/978-3-319-23132-23.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., et al. (2011). The subspace gaussian mixture model-a structured model for speech recognition. *Computer Speech and Language*, *25*(2), 404–439.

Prabhaker, M. (April 2006). Tamil market: A spoken dialog system for rural India. In ACM CHI Conference.

Rabiner, L. R. (1994). Applications of voice processing to telecommunications. *Proceedings of IEEE*, *82*, 199–228.

Rabiner, L. R. (1997). Applications of speech recognition in the area of telecommunications. IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, pp. 501–510.

Rose, Richard, & Tang Yun, (2011). An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition. In ICASSP, pp. 4508–4511.

Rose R. C, Yin, S.C., & Tang, Y, (2011). An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition, in Proc. ICASSP, pp. 4508-4511.

Sailor, H. B., & Patil, H. A. (2016). Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), 2341–2353.

Shahnawazuddin, S., Thotappa, D., Sharma, B. D., Deka, A., Parasanna, S. R. M., & Sinha, R. (2013). Assamese spoken query system to access the price of agricultural commodities. National Conference on Communications (NCC), New Delhi, India, pp. 1–5.

Thimmaraja, G. Y., & Jayanna, H. S. (2017). A spoken query system for the agricultural commodity prices and weather information access in Kannada language. *International Journal of Speech Technology (IJST)*, *20*(3), 635–644. https://doi.org/10.1007/s10772-017-9428-y.

Trihandoyo, A., Belloum, A., & Hou, K. M. (1995). A real-time speech recognition architecture for a multi-channel interactive voice response system. *Proceedings of ICASSP*, *4*, 2687–2690.

Van Segbroeck, M., & Van Hamme, H. (2011). Advances in missing feature techniques for robust large-vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 123–137.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Menlo Park: Sun Microsystems, Inc.

Wolfe, P. J., & Godsill, S. J. (Aug. 2001). Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In Proceedings of 11th IEEE Signal Process. Workshop Statist. Signal Process., pp. 496–499.

Xia, B., Liang, Y., & Bao, C. (Nov. 2009). A modified spectral subtraction method for speech enhancement based on masking property of human auditory system. International Conference on Wireless Communications Signal Processing, WCSP, pp. 1–5.

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

Zhang, S. X., Ragni, A., & Gales, M. J. F. (2010). Structured log linear models for noise robust speech recognition. *IEEE Signal Processing Letters*, *17*(11), 945–948.