Efficient Speaker Independent Isolated Speech Recognition for Tamil Language Using Wavelet Denoising and Hidden Markov Model

C. Vimala and V. Radha

Abstract Current research on Automatic Speech Recognition (ASR) focuses on developing systems that would be much more robust against variability in environment, utterance, speaker and language. In this paper all these major factors are considered to develop a system which works powerfully for recognizing a set of Tamil spoken words from a group of people at different noisy conditions. Developing an ASR system in the presence of noise critically affects the speech quality, intelligibility, and recognition rate of the system. Thus, to make a system robust against different noisy conditions, the most popular speech enhancement techniques such as spectral subtraction, adaptive filters and wavelet denoising are implemented at four SNR dB levels namely -10, -5, 5 and 10 with three types of noise such as white, pink and babble noise. This research work is carried out for developing a speaker independent isolated speech recognition system for Tamil language using Hidden Markov Model (HMM) under the above noise conditions. Better improvements are obtained when the proposed system is combined with speech enhancement preprocessor. Based on the experiments 88, 84 and 96 % of recognition accuracy are obtained from enhanced speech using Nonlinear Spectral Subtraction, RLS adaptive Filter and Wavelet approach respectively.

Keywords Nonlinear spectral subtraction · RLS adaptive algorithm · Wavelet denoising · MFCC · HMM · Tamil speech recognition

C. Vimala (✉) · V. Radha
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India
e-mail: vimalac.au@gmail.com

V. Radha
e-mail: radhasrimail@gmail.com

Mohan S. and S. S. Kumar (eds.), *Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012)*, Lecture Notes in Electrical Engineering 222,
DOI: 10.1007/978-81-322-1000-9_52, © Springer India 2013

1 Introduction

Speech recognition allows the system to identify the words that a person speaks into a microphone or telephone and convert them into written text. The biggest advantage of using ASR is the ability to achieve hands-free computing. It also offers huge benefit for people with disability who find difficulties in using a keyboard and mouse. Hence it is becoming an attractive alternate choice for users to manage applications through voice rather than a mouse or keyboard. The significant applications of ASR system include voice dialing, call routing, automatic transcriptions, information searching, data entry, speech-to-text processing and aircraft etc. Based on the requirement, the ASR system can be classified into different broad categories. Accordingly it can be classified as isolated or continuous speech and speaker-dependent or speaker-independent system. Some of the speech recognition applications require speaker-dependent system whereas some systems need speaker independent system where the inter-speaker variability should be eliminated. Based on these constraints, it is desirable in many applications requiring small well defined vocabularies which efficiently work for a group of people. Hence, today's researches are more focused on developing speaker independent isolated speech recognition systems. But, the great disadvantage of speaker-independent system is that the error rate is normally higher than speaker-dependent systems. In recent times, with the combination of more sophisticated techniques and improved independent speech recognition engines it offers excellent performance with improved productivity.

Another important aspect in speech recognition system is its performance level in noisy environment. Most of the systems achieve reliable performance in noise free environments but works very poor in noisy conditions. Developing a highly effective speech recognition system which achieves greater accuracy in noisy conditions is a challenging task [1]. Also, the greater part of these speech recognizers have been primarily developed for English language. Hence, today numerous researchers are mainly focusing on developing speech recognizers for their native languages.

The paper is organized as follows. The overview of a proposed system is given in Sect. 2 and the analysis of various speech enhancement techniques are explained in Sect. 3. The Sect. 4 deals with feature extraction using MFCC and Sect. 5 briefly explains about the HMM based speech recognition. The experimental results are shown in Sect. 6. The subjective and objective performance evaluation is given in Sect. 7. Finally, the optimum method for speaker independent isolated speech recognition system for Tamil language is concluded in Sect. 8.

2 Overview of the Proposed System

The proposed system involves various preprocessing and feature extraction techniques as front end for the ASR system. The Fig. 1 shows the overview of speaker independent isolated speech recognition system for Tamil language.

Initially, the system receives the analog signal which is converted into a digital signal using Digital Signal Processor (DSP). Next, the digitized speech is given to the speech preprocessing system which will perform dc offset removal and pre-emphasis. Subsequently, the preprocessed signal is given to the speech enhancement system for speech noise cancellation. After that, the useful feature vectors are extracted from the enhanced speech signals using MFCC. Finally, the HMM model is used to recognize the spoken word based on these feature vectors. These techniques are explained in the subsequent sections.

3 Speech Enhancement Techniques

In real time environment, speech signals are normally corrupted by numerous types of noise. The occurrence of noise in speech significantly decreases the intelligibility and the quality of the signal. Reducing noise and enhancing the perceptual quality and intelligibility of a speech without disturbing the signal quality are the vital task. Hence, speech enhancement algorithm plays a crucial role in speech recognition to improve the accuracy in noisy environment. Several techniques have been proposed for this purpose namely spectral subtraction, adaptive noise cancellation, kalman filtering, fuzzy algorithms, extended and iterative wiener filtering, HMM-based algorithms and signal subspace methods

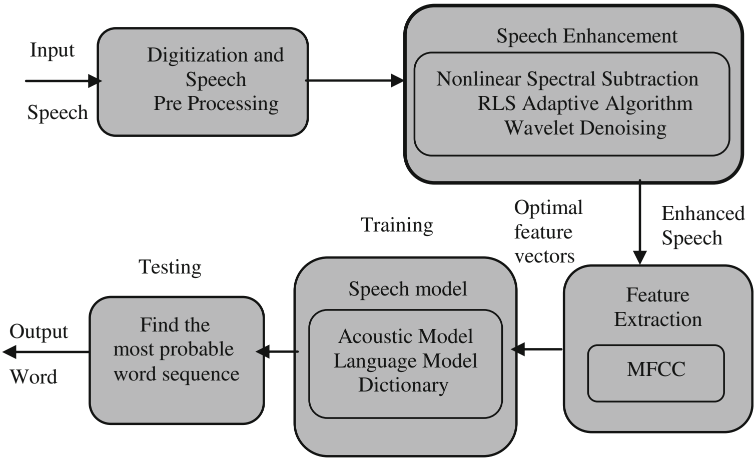


Fig. 1 Overview of Speaker independent isolated speech recognition system

etc. Achieving the high quality and intelligibility of the processed speech signal is the major target of these techniques. The main objective of speech enhancement is to improve the following:

- speech signal quality and intelligibility
- robustness of speech which is affected by noise
- speech signal-to noise ratio
- accuracy of speech recognition systems operating in noisy environments

In this paper, three types of speech enhancement techniques are implemented namely spectral subtraction, adaptive filtering and wavelet. For this work, three types of noises are considered namely white, pink and babble noise at four SNR dB levels namely −10, −5, 5 and 10. The following section gives a brief outline of these techniques.

3.1 Spectral Subtraction

The spectral subtraction is the conventional method proposed for additive background noise. It is mainly used to suppress the noise from the degraded signal and it became more popular due to its simplicity [2]. It is represented in Eq. (1)

$$y(n) = s(n) + d(n) \tag{1}$$

where y(n) is the noisy speech which is composed of the clean speech signal s(n) and the additive noise signal d(n). It is implemented by estimating the spectral magnitude during no speech and subtracting this spectral estimate of the noise from the subsequent speech spectral magnitude [3]. This technique is effective for additive noise reduction, however they introduce some artificial noise which alters the original signal. Hence these algorithms has gone through many modifications with time and introduced new methods. They include Nonlinear Spectral Subtraction (NSS), MultiBand Spectral Subtraction, Minimum Mean Square Error (MMSE), and Log Spectral MMSE [4]. In this paper, all these techniques are implemented and the performances are measured both subjectively and objectively. Based on the analysis, it was found that, the NSS is the premium method for better noise cancellation [5].

3.1.1 Nonlinear Spectral Subtraction

Among the spectral subtraction techniques, the non-linear approach gives better result since it make use of frequency dependent subtraction factor for different types of noise [6]. In this method, in order to produce improved results, larger values are subtracted at frequencies with low SNR levels and smaller values are subtracted at frequencies with high SNR levels. It is estimated using Eq. (2).

$$\begin{aligned} |X_e(\omega)| &= |Y(\omega)| - \alpha(\omega)N(\omega) \\ &\quad if \\ |Y(\omega)| &> \alpha(\omega)N(\omega) + \beta|D_e(\omega)| \\ &\quad \text{else} \\ \beta &= |Y(\omega)| \end{aligned} \tag{2}$$

where β is the spectral floor, $\alpha(\omega)$ is a frequency dependent subtraction factor and N(ω) is a non-linear function of the noise spectrum

$$N(\omega) = Max(|D_e(\omega)|) \tag{3}$$

where $N(\omega)$ is the maximum of the noise magnitude spectra.

$$\begin{aligned} \alpha(\omega) &= \frac{1}{r + P(\omega)} \\ \mathrm{y(t)} &= \mathrm{x(t)} + \mathrm{n(t)} \\ \mathrm{Y_{j,k}} &= \mathrm{X_{j,k}} + \mathrm{N_{j,k}} \end{aligned} \tag{4}$$

where r is a scaling factor and $P(\omega)$ is the square root of the posteriori SNR estimate given as

$$P(\omega) = |Y(\omega)|/|D_e(\omega)| \tag{5}$$

The performance of this algorithm is further considered for comparing with the other two enhancement techniques.

3.2 Adaptive Filtering

Adaptive filters also called self learning filters which do not have constant filter coefficients and do not require a priori knowledge of signal or noise characteristics [7]. It has the potential to achieve enhanced performance in an environment where information of the relevant statistics is not presented [8]. The well known and popular kind of adaptive filters are Least Mean Square (LMS), Normalized Least Mean Square (NLMS) and Recursive Least Squares (RLS) algorithms. The performances of all these algorithms are analyzed to find out the efficient adaptive algorithm for speech enhancement. Among these, LMS and NLMS algorithms are very simple and effective method to implement but they are slower. Whereas the RLS algorithm makes the converging speed and also offers better noise reduction and enhanced speech quality and intelligibility when compared to the other algorithms. The RLS algorithm and its advantages are discussed below.

3.2.1 Recursive Least Squares Algorithm

RLS algorithms offer excellent performance in time varying environments like speech [9]. It is a recursive implementation of the Wiener filter which is used to find the filter coefficients that relate to producing the recursively least squares of the error signal i.e. the difference between the desired and actual signal. In contrast to LMS and NLMS algorithm, the RLS aims to reduce the mean square error. At each instant, an exact minimization of the sum of the squares of the desired signal estimation errors are performed by referring the values of previous error estimations. The RLS approach offers faster convergence and smaller error with respect to the unknown system, at the expense of requiring more computations [10]. In this paper, the performance of LMS, NLMS and RLS algorithms are experienced at different noisy conditions. As a result, it was observed that the performance of RLS algorithm is superior to other adaptive algorithms [10].

3.3 Wavelet Denoising

Wavelet denoising is a non-parametric estimation method that has been proposed in recent years for speech enhancement applications. Transform domain always plays an important role in any speech signal processing application. Fourier transform was the earlier choice of domain but creates annoying musical noise in speech noise suppression [11]. Later, some methods have been proposed to solve this problem but have not achieved satisfied performance. In recent years, wavelet domain based approach has been found to be a very useful tool for solving various problems particularly for speech denoising [11]. Compared to Fourier transform, it is possible in wavelets to obtain a good approximation of the given function *f* by using only a few coefficients which is the great metric [12]. If an observed signal includes unwanted noise, the result is an additive signal model given by (6)

$$\mathrm{y(t)} = \mathrm{x(t)} + \mathrm{n(t)} \tag{6}$$

where ***y*** is the noisy signal, ***x*** is the clean signal, and ***n*** is the additive noise.

Then the wavelet transform performs the following (7)

$$\mathrm{Y_{j,k}} = \mathrm{X_{j,k}} + \mathrm{N_{j,k}} \tag{7}$$

where $\mathrm{Y_{j,k}}$ represents the kth set of wavelet coefficients across the selected scale j.

In wavelet transform, the noise is typically represented across time with smaller coefficients whereas signal energy is focused on larger coefficients. This improves the possibility of separating the signal from the noise based on some threshold [12]. The main advantage of wavelet denoising is that it removes noise from the corrupted signal without modifying the speech content. Among different types of wavelet family, Daubechies wavelets are a powerful and efficient approach as they are orthogonal and localized both in real and Fourier space [12]. In this paper, level 8 Daubechies 4 wavelet has been implemented which offers comparatively good performance than some other wavelets family.

All the above three algorithms are implemented and their performances are measured both subjectively and objectively. For the experiments, the separate noise corpus from NOIZEUS were collected and added to the Tamil Speech signals. Analysis is done on noisy speech signal corrupted by white, pink and babble noise at −10, −5, 5 and 10 SNR dB levels. The following Fig. 2 shows the results of speech enhancement using Nonlinear Spectral Subtraction, RLS adaptive algorithm and Wavelet approach for a sample signal corrupted by 10 dB babble noise. It is clear from the experiments that the wavelet technique offers better performance for all types of noise and SNR levels which are considered for this work. The subjective measures of these techniques are explained in the performance evaluation section. The enhanced output signals from these techniques are taken as an optimum input signal for feature extraction.

4 Speech Feature Extraction

The most significant part of all recognition systems is the feature extraction, which converts the speech signal into some digital form of meaningful features.

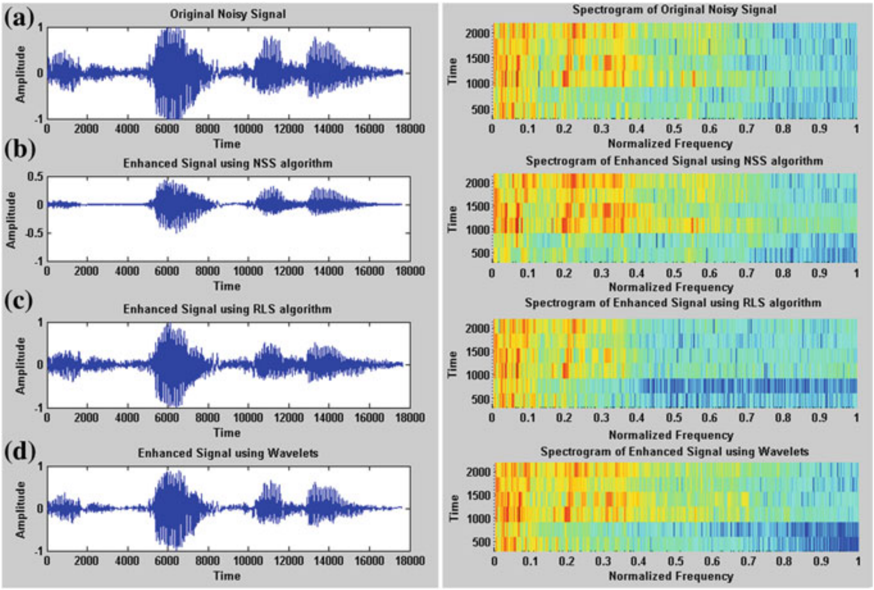


Fig. 2 Results of speech enhancement and its spectrogram representations. **a** Original babble noise signal at 10 dB SNR. **b** Enhanced signal using nonlinear spectral subtraction. **c** Enhanced signal using RLS adaptive algorithm. **d** Enhanced signal using wavelet

Providing prominent features is the major task in speech recognition system to achieve good accuracy. The good choice of features makes the speech recognition job easier in classification. The widely used feature extraction techniques are Linear Predicting Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC), RASTA and Perceptual Linear Prediction (PLP). In this paper, the MFCC feature extraction technique is chosen due to its prominent characteristics. The following section gives details about MFCC feature extraction.

4.1 Mel Frequency Cepstral Coefficients

The human ear has high frequency resolution in low-frequency parts of the spectrum and low frequency resolution in the high-frequency parts of the spectrum. The coefficients of the power spectrum of a speech signal can be transformed to reflect the frequency resolution of the human ear. MFCC is based on the short-term analysis, and thus from each frame a feature vector is computed. To extract these coefficients the speech sample is taken as an input and hamming window is applied to minimize the discontinuities of a signal. Then Discrete Fourier Transformation (DFT) is used to generate the Mel filter bank. According to Mel frequency wrapping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping this log energy, the numbers of coefficients are obtained. This coefficient takes the best effects of Discrete Cosine Transformation (DCT) for the cepstral coefficients calculation. MFCC can be computed by using the Eq. (8)

$$\mathrm{Mel}\ (\mathrm{f}) = 2595^{*}\log 10\,(1 + \mathrm{f}/700) \tag{8}$$

Typical feature vector of 39 different parameters for each 10 ms of speech are extracted. It can be observed from the experiments that the performance of the system can be improved if the energy and delta coefficients are used additionally. These features are given as the input for the post processing steps and they are explained in the following section.

5 Speech Recognition Using HMM

Undoubtedly, modern general-purpose speech recognition systems are based on HMM. The reason for using HMM is that it can be trained automatically and it is simple and computationally feasible to use. The greatest advantage of HMM model is it is more robust to environment noise and distortions. It is a stochastic approach with set of states and transition probabilities between those states. Here, each state describes a stationary stochastic process and the changeover from one state to another to define how the process changes its characteristics in time [13].

Each state of the HMM can model the generation of the observed symbols using a stationary stochastic emission process [13]. It allows phoneme or word rather than frame-by-frame modeling of speech. Basically, the problem of speech recognition can be stated as follows. In a given acoustic observation X = X_1, X_2…X_n, the objective is to find out the matching word sequence W = W_1, W_2…W_m which has the maximum posterior probability P(W|X) [14]. It can be defined by the Eq. (9)

$$W = \arg\max_{w} P(W/X) = \arg\max_{w} \frac{P(W)P(X/W)}{P(X)} \tag{9}$$

where P(W) is the probability of word W uttered and P(X|W) is the probability of acoustic observation of X when the word W is uttered. P (X|W) is also known as class conditioned probability distribution. P(X) is the average probability that observation X will occur [14]. The goal is to find the word W by maximizing the probability of X. To accomplish the above task an ASR system requires three significant components which are also considered as post processing steps. They are acoustic model, dictionary and language model. Since all these models are language independent, these models are created manually for Tamil language. Once all these models are developed, then the HMM model will perform likelihood evaluation, state sequence decoding and HMM estimation to find out the most probable word sequence from the given observation vector. The experimental results and the performance evaluation of the proposed system are given in the following sections.

6 Experimental Results

The main objective of this research work is to develop an ASR system to recognize isolated speech in Tamil Language from a group of speakers. Speaker independent speech recognition system itself makes the system complex. In addition to speaker characteristics, the environment, vocabulary size, Signals to Noise Ratio (SNR) also make the system very complex to implement. In this paper, all these factors are considered but the vocabulary size is limited to 50 words. Since speech corpora are not available for Tamil language it is created manually. The corpus containing 50 utterance of isolated speech were collected from 10 females. The database consists of 10 repetitions of every word produced by each speaker. The experiments were carried out in three conditions namely clean speech, noisy speech and enhanced speech with different type of noisy conditions. The best input signal for speech recognition system is identified according to the results obtained from speech enhancement. Next, the useful feature vectors extracted using MFCC are given as the input for the HMM based speech recognition. It can be observed that the HMM offers better results and higher accuracy in clean and enhanced signals. The wavelet based enhanced signals has achieved a great performance compared to Nonlinear Spectral Subtraction and RLS adaptive filters. In this work, both objective and subjective measures are considered to ensure the performance of speech enhancement techniques for improved accuracy and they are explained in the next section.

7 Performance Evaluations

Developing an ASR system continues to be a challenging field for researchers due to a great number of factors, which cause variability in speech. The main objective of this work is to recognize a set of Tamil words from a group of people under different noisy conditions. In this paper, the performance evaluations are done for both speech enhancement and recognition separately.

7.1 Evaluation of Speech Enhancement Techniques

For speech enhancement, three types of techniques are employed namely Non-linear Spectral Subtraction, RLS adaptive algorithm and wavelet based denoising approach. These algorithms are measured based on three metrics namely Mean Square Error (MSE) Rate, Peak Signal to Noise Ratio (PSNR) and Signal to Noise Ratio (SNR). Based on these measures it was found that the wavelet based denoising method confirms its superiority by offering less MSE and higher SNR and PSNR values. The wavelet method has produced better signal quality and intelligibility in all the noisy conditions and at different SNR dB levels which are implemented in this research work. The performance of wavelet was found to be comparatively good with subjective listening tests also.

7.2 Evaluation of Speech Recognition Accuracy

The speech recognition system is generally measured in terms of Word Error Rate (WER), which is the ratio between misclassified words, and total number of tested words. Generally, ASR research has been focused on minimizing the recognition error to zero in real-time independent of vocabulary size, noise, speaker characteristics and accent. The proposed system can able to recognize 48/50 words on an average by using clean signal. The following Figs. 3 and 4 illustrate the performance of the proposed system based on accuracy rate and WER for different types of signal employed in this research work.

The average accuracy of a given database is shown in Fig. 3 and it is clear from the above figures that the wavelet approach with HMM performs extremely well for all the datasets at different conditions for all the speakers enrolled in the database. The results prove that the proposed system offered higher accuracy level and less WER.

Fig. 3 Performance evaluation of proposed system based on accuracy

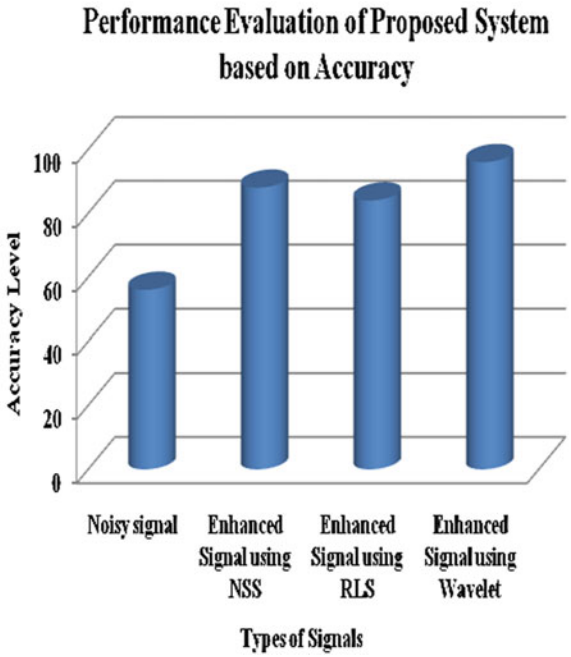
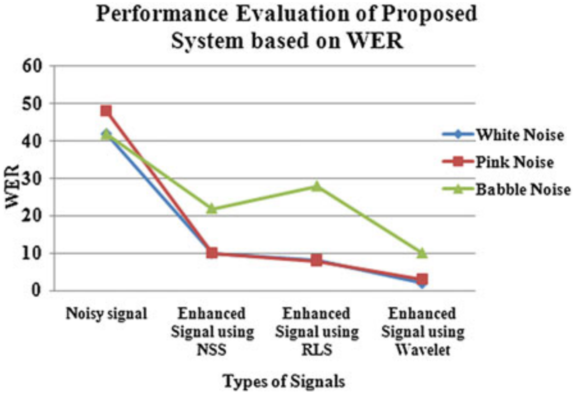


Fig. 4 Performance evaluation of proposed system based on WER



8 Conclusion

In recent years, considerable amount of research works were focused on human computer interaction through speech recognition system. Whereas language barrier is one of the significant factors which make these technologies less accessible. Based on this specific need, in this paper an efficient speaker independent isolated speech recognition system is implemented for Tamil language. The proposed system is developed using HMM since it is a most flexible and successful approach to speech recognition. The main objective of this work is to recognize the spoken words under different noisy conditions. In this paper, three types of noises are used namely white, pink and babble noise at four SNR dB levels like −10, −5, 5 and 10. The most popular speech enhancement techniques namely spectral subtraction, adaptive filters and wavelet denoising are implemented for removing the above mentioned noise types. The performances of these techniques were evaluated based on MSE, PNSR and SNR values. Based on the results, it was found that the spectral subtraction algorithm has achieved good noise cancellation but fails to produce intelligibility in enhanced speech. The RLS adaptive algorithm also performed well and offered good results in speech quality and intelligibility to some extent. Compared to the above algorithms, very good performance was achieved through wavelet denoising for all the noise factors considered for this work. The prominent input obtained from these speech enhancement methods are further analyzed with speech recognition accuracy. The comparison is done with different conditions like clean, noisy and enhanced signals. Comparatively, excellent accuracy and minimum WER are obtained through wavelet approaches. The system is found to be successful as it can identify spoken word even in noisy conditions which is the greatest advantage of HMM. Based on the experiments, it is concluded that the combination of wavelet and HMM approach can yield a highly effective recognition performance for speaker independent isolated speech recognition for Tamil language.

References

1. Baker JM, Deng L, Khudanpur S, Lee C-H, Glass J, Morgan N (2006–2007) Historical development and future directions in speech recognition and understanding. MINDS. Report of the speech understanding working group
2. Krishnamoorthy P, Mahadeva Prasanna SR (2009) Temporal and spectral processing methods for processing of degraded speech: a review. IETE Tech Rev 26(2):137–148
3. Fukane AR, Sahare SL (2011) Different approaches of spectral subtraction method for enhancing the speech signal in noisy environments. Int J Sci Eng Res 2(5). ISSN 2229-5518
4. Goel1 P, Garg A (2012) Developments in spectral subtraction for speech enhancement. Int J Eng Res Appl (IJERA). 2(1):055–063. ISSN: 2248-9622
5. Vimala C, Radha V (2012) A family of spectral subtraction algorithms for tamil speech enhancement. Int J Soft Comput Eng (IJSCE) 2(1). ISSN: 2231-2307
6. Lockwood P, Boudy J (1992) Experiments with a nonlinear spectral subtractor (NSS) hidden markov models and the projection for robust speech recognition in cars. Speech Commun 11(2–3):215–228
7. JaganNaveen V, Prabakar T, Venkata Suman J, Devi Pradeep P (2010) Noise suppression in speech signals using adaptive algorithms. Int J Signal Process Image Process Pattern Recogn 3(3):87–96
8. Hadei SA, Student member IEEE, Lotfizad M (2010) A family of adaptive filter algorithms in noise cancellation for speech enhancement. Int J Comput Electr Eng 2(2):1793–8163
9. Borisagar KR, Kulkarni GR (2010) Simulation and comparative analysis of LMS and RLS algorithms using real time speech input signal. Global J Res Eng 10(5):44 (Ver1.0)

10. Vimala C, Radha V (2012) Optimal adaptive filtering technique for tamil speech enhancement. Int J Comput Appl (0975–8887) 41(17):23–29
11. Chavan MS, Chavan MN, Gaikwad MS (2010) Studies on implementation of wavelet for denoising speech signal. Int J Comput Appl (0975–8887) 3(2):1–7
12. Johnson MT, Yuan X, Ren Y (2007) Speech signal enhancement through adaptive wavelet thresholding. Speech Commun 49(2):123–133
13. Lama P, Namburu M (2010) Speech recognition with dynamic time warping using MATLAB. CS 525, SPRING 2010—Project report
14. Thangarajan R, Natarajan AM, Selvam M (2008) Word and triphone based approaches in continuous speech recognition for tamil language. WSEAS Trans Signal Process 4(3). ISSN: 1790-5022