Speaker Independent Connected Digit Recognition Using VQ and HMM in Additive Noise Environment

A. Revathi* and Y. Venkataramani**

Saranathan College of Engineering, Trichy
revathidhanabal@rediffmail.com, principal@saranathan.ac.in

Abstract. The main objective of this paper is to discuss the effectiveness of concatenated perceptual features and the noise reduction technique based on wavelet transform and Recursive least square filtering in getting the good recognition rate for the peculiar combination of connected digits in additive noise environment. The proposed concatenated perceptual features are captured and code book indices are extracted. Expectation maximization algorithm is used to generate discrete HMM models for the connected digits. Speech recognition system is evaluated on clean and noisy test speeches and the selection is based on which model gives maximum log likelihood value. Speeches for this work are randomly chosen from "TI Digits_1", "TI Digits_2" databases. This concatenated perceptual feature yields the accuracy of 81.4% and 73% for the combination of connected digits (10 – 19) and (12-19,21,31,41,51,61,71,81,91). Pink noise, white noise, babble noise and factory noise are considered in this work.

Keywords: Hidden markov model (HMM), Frequency response, Speech recognition, Vector quantization (VQ), Perceptual linear predictive cepstrum (PLP), Noise, Wavelet transform, Recursive least square (RLS) filtering.

1 Introduction

Speech recognition involves the decoding of speech signal in sequential manner based on the observed acoustic features of the signal and exploitation of known relations between acoustic features and phonetic symbols. Evaluation of speech recognition system on clean training and test speeches normally provides good accuracy. Nowadays, it becomes a challenging task to provide a robust speech recognition system in the presence of stationary and non-stationary noise. Our goal is to maximize the speech recognition rate in a noisy environment. Possible applications of this work are recognition of telephone numbers or bank account numbers from a noisy recorded speech by intelligence and police surveillance, investigation of disputed credit card number provided over phone in a noisy environment. Yuval Cohen et.al [1] discussed

* Professor, Dept.of ECE.
** Director.

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 393–402, 2012.

the application of speech enhancement algorithm to evaluate the connected word recognition in a noisy environment. Mosakiyo Fujimoto et.al [2] used GMM based speech estimation method and EM based noise estimation method are used for evaluating the noisy speech recognition systems. Carlos Lima et.al has done spectral normalization [3] to improve the accuracy of isolated word recognition. Synaptic adaptation and two tone suppression techniques are implemented by Serajul Haque et.al [4] to enhance speech recognition accuracy. Multi band approach using wavelet transform is used by Wesam Alkhaldi et.al [5] in speech recognition system. Throat microphone for accurate voicing detection is used by Tomas Tebem et.al [6] to improve the performance of the speech recognition system. MFCC and wavelet packets are used by Phani Kumar et.al [7] for speech recognition in a noisy environment. Pitch detection approach is used by Rashmi Makhaijani et.al [8] to enhance speech in speech recognition. Akshiyo K.Swain et.al [9] extracted unbiased LPC parameters by using orthogonal least squares method to improve the speech recognition rate. LMS adaptive filters are used by Jose Louis Oropezo Radriguez et.al [10] to improve the speech recognition rate in noisy environments. LMS adaptive filters and wavelets are used by Jose Louis Oropezo Radriguez et.al [11] to improve the speech recognition rate in noisy environments. A.Revathi et.al [12] anaysed the use of perceptual features and iterative clustering approach for performing isolated digits/continuous speech recognition. Combination of vector quantization and HMM is used by A.Revathi et.al [13] to evaluate the speech recognition system. Voice activity detection algorithm is used by Xiaokun Li et.al [14] to improve the recognition performance in noisy environments. Statistical model based voice activity detection and noise suppression is used as front end tool by Mosakiyo Fujimoto et.al [15] for automatic speech recognition in noisy environments. Syllables are used as acoustic units by Azmi M.M. et.al [16] to perform Arabic speech recognition in noisy environment. In this work, stationary and non-stationary noises such as white noise, factory noise, pink noise and babble noise are considered to evaluate the noisy speech recognition system. Better results are due to the implementation of additional preprocessing techniques such as wavelets and adaptive RLS filters prior to conventional preprocessing stages. Results obtained are comparable to the case of clean speech recognition.

2 Feature Based on Cepstrum

The short-time speech spectrum for voiced speech sound has two components: 1) harmonic peaks due to the periodicity of voiced speech 2) glottal pulse shape. The excitation source decides the periodicity of voiced speech. It reflects the characteristics of speaker. The spectral envelope is shaped by formants which reflect the resonances of vocal tract. The variations among speakers are indicated by formant locations and bandwidth.

2.1 PLP Extraction

PLP (perceptual linear predictive cepstrum) speech analysis method [17-19] models the speech auditory spectrum by the spectrum of low order all pole model. The detailed procedure for PLP (perceptual linear predictive cepstrum) extraction is given below. This perceptual feature mainly emphasizes the need for critical band analysis which integrates energy spectral density in the frequency range (0-8) kHz for obtaining the speech auditory spectrum. Loudness equalization is a pre emphasis block used to emphasize the upper and middle part of the spectrum and cube root compression is done to reduce the dynamics of the speech spectrum. Block diagram for perceptual features extraction is shown in Fig.1.The relationship between frequency in Bark and frequency in Hz is specified as in (1)

$$f(bark) = 6 * arcsinh(f(Hz)/600) \tag{1}$$

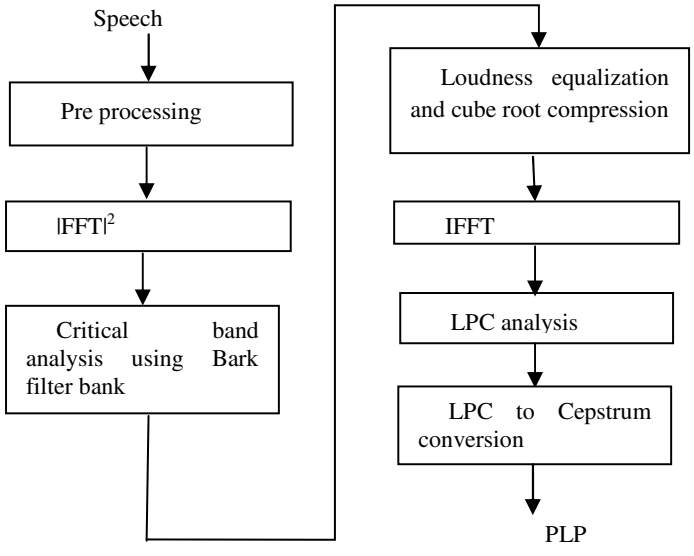

Fig. 1. PLP Extraction Model

3 Speech Recognition Based on VQ and HMM

Speech database considered in this work contains speeches of 8 female speakers and 8 male speakers**.** 8 speeches of 4 female and 4 male speakers are used for training. 10 speeches of other 4 female and 4 male speakers are used for testing. Connected digit recognition system is evaluated for 800 test speeches for connected digits (10-19) and 1280 test speeches for connected digits (12-19,21,31,41,51,61,71,81,91). Connected

digits are formed by concatenating the respective isolated digits from TI digits_1 and TI digits_2 database. For creating a training model, speech signal is first pre-emphasized using a difference operator. Hamming window is applied on differenced speech frames of 16 msecs duration with overlapping of 8 msecs. Then the PLP feature is extracted and it is concatenated with its differential feature. For each training model corresponding to connected digits, training set of K utterances are used (spoken by many speakers) where each utterance constitutes an observation sequence of some appropriate spectral or temporal representation. For each digit or speech, HMM models are developed with state transition probability distribution, observation symbol probability distribution and initial probability distribution that optimize the likelihood of the training set observation vectors. For discrete HMM, models are initialized with 256 observation sequences and 8 states. Code book indices are used as input to train the models. Clustering algorithm is used to obtain cluster centers and code book indices for training vectors of the connected digits considered for training.To evaluate the performance of the noisy speech recognition system, noises such as babble noise, white noise, factory noise and pink noise are taken from "Noise data" database. Noises are added to the test speeches at various levels so that SNR values of different ranges are obtained. Speech recognition is done by using the combination of VQ and discrete HMM techniques [20,21]. For discrete HMM, models are initialized with 256 observation sequences and 8 states and code book indices are used as input to train the models. In some cases, noise is added to the test speech in such a way that SNR is found to be negative (i.e.) noise energy is dominating the speech energy. In this work, noise suppression technique based on wavelets and adaptive RLS filtering as additional preprocessing technique is implemented along with conventional preprocessing stages such as pre emphasis, frame blocking and widowing to enhance the performance of the system. This noisy test speech undergoes first level of wavelet decomposition using 'Haar' wavelets and the approximation coefficients in the first level are suitably up sampled to generate a signal in which there is no contribution due to high frequency disturbance. Adaptive RLS filtering is subsequently applied and the filter coefficients are adaptively changed to minimize the least square error between the desired output and the actual output. Finally, noise reduced speech similar as that of clean test speech is obtained.

For testing, observation sequences of the feature vectors of the clean or noisy test speeches are applied to all the training HMM models. HMM models for each connected digit are already trained with state probability distribution matrices, observation symbol probability distribution matrices and initial probability distribution vectors. For each model, log likelihood values are calculated. Selection of the speech or digit is done by comparing likelihood values and recognized speech is the one whose model likelihood is the highest.

4 Results and Discussion

The performance of clean connected digits recognition system based on concatenated perceptual features is evaluated by applying test speech vectors to all the HMM training models. Log likelihood values for each model are computed. Selection is

based on the comparison of log likelihood values and decision is made with respect to the model which provides the maximum log likelihood value. Speech recognition rate is the number of correct choices over the total number of test speeches. System is evaluated for 80 test speeches for each connected digit. The individual accuracy of peculiar combination of connected digits (12,13,14,15,16,17,18,19,21,31,41,51,61,71,81,91) is shown as a bar chart in Fig.2 and there is a clear indication of obtaining 100% accuracy for one connected digit (31). This connected digit is considered for the evaluation of noisy speech recognition system for some of the low and high frequency noises whose frequency distribution characteristics are depicted in Fig.3.

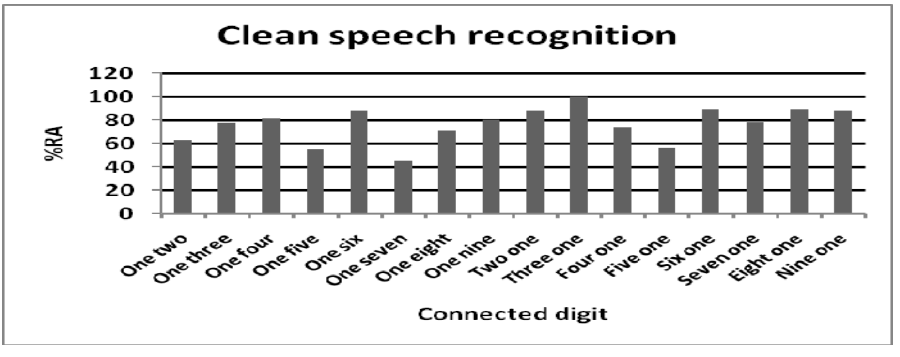

Fig. 2. Comparison chart of individual accuracy of connected digits (12,13,14,15,16,17,18, 19,21,31,41,51,61,71,81,91)

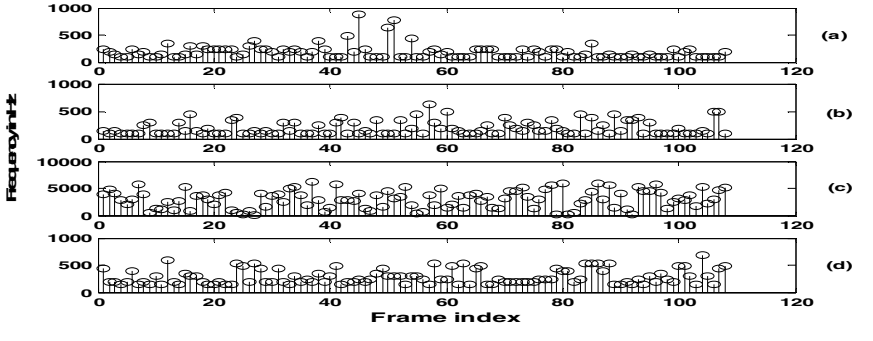

Fig. 3. Illusration of the frequency distribution of noises (a) Factory noise (b) Pink noise (c) White noise (d) Babble noise

From the Fig.3, it is evident that factory noise, pink noise and babble are low frequency noises and white noise is a high frequency noise. Following plots in Fig.4 and Fig.5 indicate the effectiveness of the additional preprocessing technique based on wavelets and adaptive RLS filtering in removing additive low frequency babble noise (SNR = 0db) from noisy speeches.

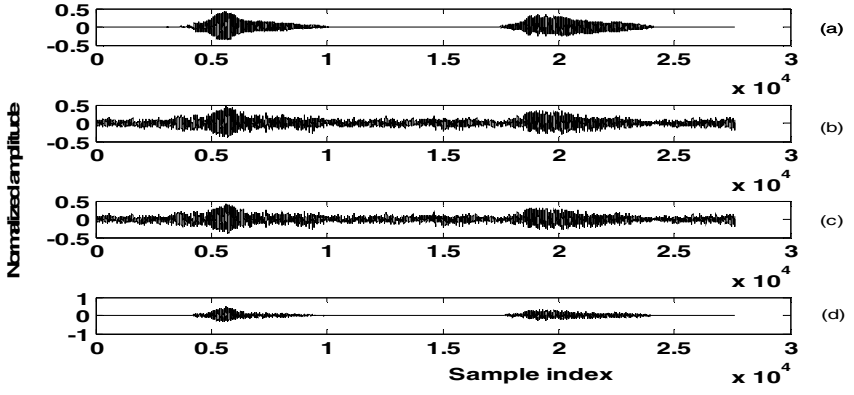

Fig. 4. Illusration of the effect of addional preprocessing for babble noise using signal characteristics. (a) Clean speech (b) Noisy speech (c) reconstructed speech after first level of wavelet decomposition (d) Noise reduced speech after RLS filtering.

Spectrogram plots in Fig.5 depict the importance of the additional preprocessing techniques with respect to the retention of speech frequencies in the noise reduced speech after RLS filtering stage.

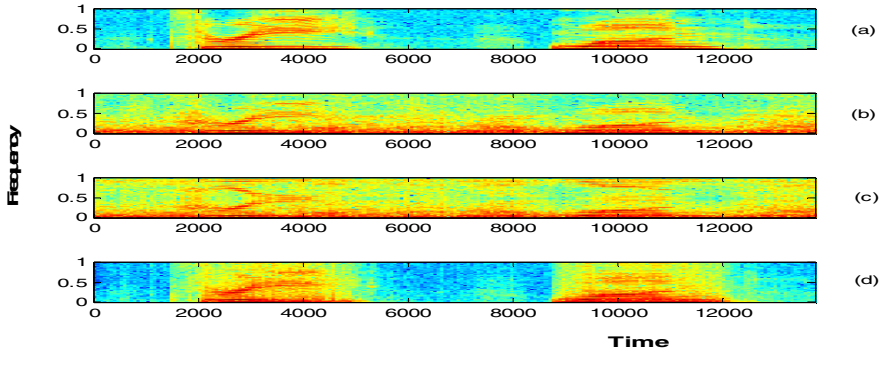

Fig. 5. Illusration of the effect of addional preprocessing for babble noise using spectrogram. (a) Clean speech (b) Noise speech (c) reconstructed speech after first level of wavelet decomposition (d) Noise reduced speech after RLS filtering.

Following plots in Fig.6 and Fig.7 indicate the effectiveness of the additional preprocessing technique based on wavelets and adaptive RLS filtering in removing additive high frequency white noise (SNR = 7db) from noisy speeches.

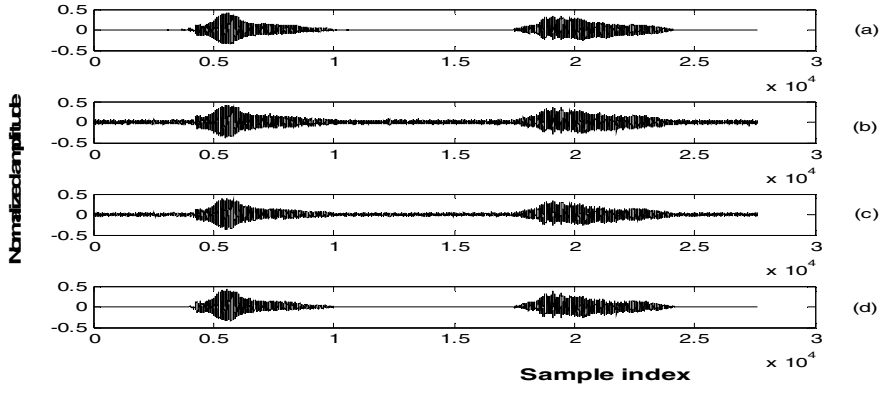

Fig. 6. Illusration of the effect of addional preprocessing for white noise using signal characteristics. (a) Clean speech (b) Noisy speech (c) reconstructed speech after first level of wavelet decomposition (d) Noise reduced speech after RLS filtering.

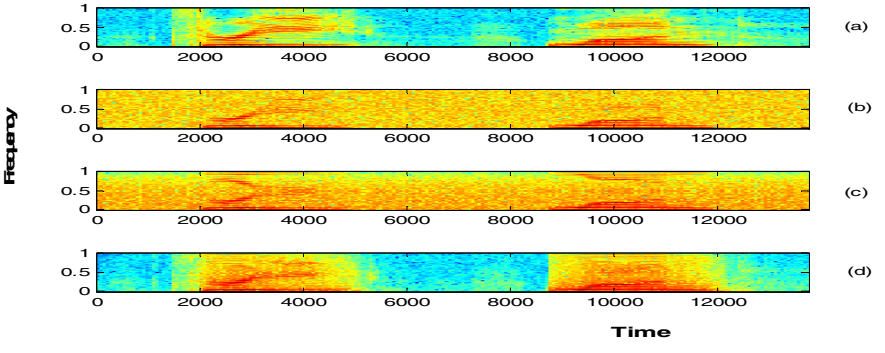

Fig. 7. Illusration of the effect of addional preprocessing for white noise using spectrogram. (a) Clean speech (b) Noisy speech (c) reconstructed speech after first level of wavelet decomposition (d) Noise reduced speech after RLS filtering.

Fig.8 demonstrates the frequency distribution of the clean speech, noisy speech (Addition of pink noise to the test speech at SNR = 1 db), Reconstructed speech after wavelet decomposition and noise reduced speech after RLS filtering stage. From these plots, it is evident that most of the low frequencies present in the clean speech are reproduced in the noise reduced speech. Subjective test is performed on the noise reduced speeches and it is clear that the noise reduced speeches utter in the similar manner as that of the clean test speeches.

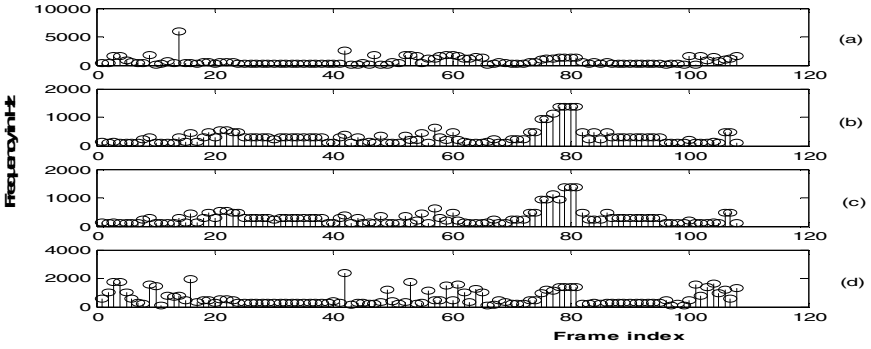

Fig. 8. Illusration of the effect of addional preprocessing for pink noise using frequency distribution characteristics. (a) Clean speech (b) Noisy speech (c) reconstructed speech after first level of wavelet decomposition (d) Noise reduced speech after RLS filtering.

Fig.9 is the comparison chart of the evaluation of the noisy speech recognition system. Pink noise, factory noise, babble noise and white noise are added to the test speeches at various levels and it is understood that there is not too much deterioration in terms of accuracy even if the noise energy is dominating the signal energy for several cases. From the Fig.9, it is understood that the system gives the better accuracy for the addition of pink noise to the test speeches at various levels.

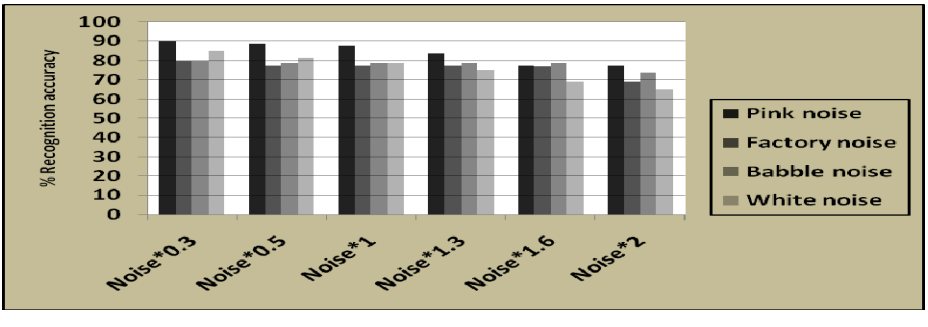

Fig. 9. Illustration of the effect of noises on accuracy

5 Conclusions

This paper proposes the use of additional preprocessing technique and concatenated PLP feature for evaluating VQ+HMM based speaker independent peculiar combination of connected digits recognition schemes and the evaluation is done on clean and noisy test speeches. In HMM based technique, discrete HMM models are developed using code book indices as input and these models developed for connected digits are considered for system evaluation. Perceptual based features normally perform well in developing robust speech recognition system, because they

inherently depict the perceptually important characteristics of the speech. Even though the noises considered in this work have frequencies falling in the speech frequency range, accuracy is not very much degraded because the noise reduced speech duplicates all the frequencies present in the clean test speech. Noise reduction is mainly performed by applying the combinational technique based on wavelet transform and adaptive RLS filtering on the noisy test speeches. Better results are obtained for noisy speech recognition even though the significant amount of noise is added to the test speeches. This is actually due to the use of additional preprocessing technique based on wavelets and RLS filtering along with conventional preprocessing. This additional preprocessing works well for both low frequency and high frequency noises.

References

[1] Cohen, Y., Erell, A.: Enhancement of connected words in an extremely noisy environments. IEEE Transactions on speech and Audio Processing 5(2), 141–148 (1997)
[2] Fujimoto, M., Ariki, Y.: Robust speech recognition in additive and channel noise environments using GMM and EM algorithm. In: Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing, pp. 941–944 (2004)
[3] Lima, C., Almeida, L.B., Mohreiro, J.L.: Robust feature extraction for speech recognition in noisy environment. In: Proc. Int. Conf. Signal Processing, vol. (6), Beijing (2002)
[4] Haque, S., Togoui, R., Zakrich, A.: Perceptual features for automatic speech recognition in noisy environments. Int. Journal on Speech Communication 1(51), 58–75 (2009)
[5] Weaam, A., Fakhir, W., Hamdy, N.: Automatic speech Recognition in noisy environments using wavelet transforms. In: Proc. 45th Midwest Symposium on Circuits and Systems, Oklahoma, pp. 463–466 (2002)
[6] Tebem, T., Verhelst, W., Capman, F., Beangenlie, F.: Improved speech recognition in noisy environments by throat microphone for voice activity detection. In: Proceedings of 18th Europeon Signal Processing Conference, pp. 1978–1982 (2010)
[7] Phani Kumar, P., Vardhan, K.S.N., Sriramakrishna, K.: Perfromance evaluation of MLP for speech recognition in noisy environments using MFCC and wavelets. Int. Journal on Computer Science and Communication 1(2), 41–45 (2010)
[8] Makhaijani, R., Shrawardkar, U., Thakre, V.M.: Speech enhancement using pitch detection approach for noisy environment. Int. Journal on Engineering science and Technology 3(2), 1764–1769 (2011)
[9] Swain, A.K., Abdullah, W.: Estimation of LPC parameters of speech signals in noisy environments. In: Proceedings of TENCON, Thailand, pp. 139–142 (2004)
[10] Rodríguez, J.L.O., Guerra, S.S., Fernández, L.P.S.: Using Adaptive Filter to Increase Automatic Speech Recognition Rate in a Digit Corpus. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 78–87. Springer, Heidelberg (2007)
[11] Rodríguez, J.L.O., Guerra, S.S.: Using Adaptive Filter and Wavelets to Increase Automatic Speech Recognition Rate in Noisy Environment. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 1015–1024. Springer, Heidelberg (2007)
[12] Revathi, A., Venkataramani, Y.: Perceptual features based isolated digit and continuous speech recognition using iterative clustering approach. In: Proc. IEEE International Conference on Networking and Communication, Chennai, pp. 155–160 (2009)

[13] Revathi, A., Venkataramani, Y.: Speaker independent continuous speech and isolated digits recognition using VQ and HMM. In: Proc. IEEE Int. Conf. on Communication and Signal Processing, Calicut, pp. 198–202 (2011)
[14] Li, X., Deng, Y.: Combining speech energy and edge information for efficient voice activity detection in noisy environments. In: Proc. 19th Int. Conf. on Pattern Recognition, Tamba, FL, pp. 1–4 (2008)
[15] Fujimoto, M., Ishizuka, K., Nakatani, T.: Study of Integration of Statistical Model-Based Voice Activity Detection and Noise Suppression. In: Proc. Annual Conf. of INTERSPEECH, Brisbane, Australia, pp. 2008–2011 (2008)
[16] Azmi, M.M., Tolba, H.: Syllable based automatic Arabic speech recognition in noisy environment. In: Proc. Int. Conf. on Audio, Language and Image Processing, Sanghai, pp. 1436–1441 (2008)
[17] Hermansky, H., Tsuga, K., Makino, S., Wakita, H.: Perceptually based processing in automatic speech recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Tokyo, vol. 11, pp. 1971–1974 (1986)
[18] Hermansky, H., Margon, N., Bayya, A., Kohn, P.: The challenge of Inverse E: The RASTA PLP method. In: Proc. Twenty Fifth IEEE Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, USA, vol. 2, pp. 800–804 (1991)
[19] Hermansky, H., Morgan, N.: RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2(4), 578–589 (1994)
[20] Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice Hall, NJ (1993)
[21] Rabiner, L., Juang, B.H.: Hidden Markov models for Speech Recognition. Proc. Technometrics 33(3), 251–272 (1991)