Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 7 (24): 3935-3942, 2007
ISSN 1812-5654

Noise Robust Isolated Word Recognition Using Speech Feature Enhancement Techniques

M. Frikha and A. Ben Hamida
Ecole Nationale d'Ingénieurs de Sfax, 'ENIS',
Department of Génie Electrique, BP W, 3038, Sfax, Tunisia

Abstract: This study examines the problem of Automatic Speech Recognition (ASR) in the presence of additive interfering noise. It investigates several noise reduction techniques which are integrated into the front end of a Hidden Markov Model (HMM) isolated word recognition in order to guarantee high performance and robust recognition system. The algorithms inherent to these techniques are studied from a theoretical view point. Their implementation is described and they are tested on the TIMIT database for an isolated word recognition task. Computer experiments were carried out on both clean and noisy words using four kinds of acoustic features. Our first experiment on clean conditions showed the best performance of static acoustic features augmented by the frame's log energy and their first derivatives coefficients. The robustness of these kinds of features was tested in the second experiment. The observed average loss of performance for the perceptually based acoustic features ranges from 15 to 65% for SNR ranging from 20 to 0 dB. In the last experiment, the evaluation of two speech enhancement techniques was performed. Results revealed the effectiveness of these two techniques in such application. In fact, a maximum relative recognition rate improvement of performance up to 35% for SNR of 0 dB is obtained and this with respect to the results obtained in the second experiment.

Key words: Automatic speech recognition, acoustic front end processing, Hidden Markov Model, spectral subtraction, discrete wavelet transform

INTRODUCTION

During the last decade, there has been much research interest in the domain of robust speech recognition to improve speech recognition systems. As recognition spoken language technologies are being transferred to real word applications, the need for greater robustness against noisy environments is becoming increasingly apparent. It is well known that the best recognition accuracies are obtained when reference and test utterances are collected under the same conditions (Frikha *et al*., 2005) and that when we add noise to speech signals, a dramatic decrease of recognition performance is observed. However, real word conditions differ from ideal or laboratory conditions, causing mismatch between training and testing and consequently, inducing performance degradation in the automatic speech recognition systems. A simple special case of mismatch situation is encountered when the testing signal is corrupted by various additive noises while the training data are clean. This mismatch is considered to be the main cause of limitation in word recognition accuracies in current state of the art speech recognition systems.

There are many levels at which improvements can be made to system robustness. The process of providing robustness to the recognizer can be accomplished in three different stages: (i) the parametric stage, by means of parametric representations of speech characteristics which may show immunity to the noise process (Hegde, 2005), (ii) the acoustical stage giving rise to speech enhancement techniques (Choi, 2004; Cui and Alwan, 2005) and (iii) the modeling stage, combining adequate models of noise and clean signal in order to recognize noisy speech (Zhang and Furui, 2004).

This study is primarily intended with the two first approaches. We started evaluating the performance of four types of acoustic features (MFCC, PLP, LPC and LPCC) for clean and additive noisy speech and compared their performances. And then, we addressed the problem of enhancing speech features, which has been degraded by additive noise, before they are fed to the recognizer. We've been interested particularly in the uncorrelated additive noise because it is frequent in many real life situations and a great attention has been devoted to reduce the distortion introduced by such type of noise.

Corresponding Author: Mondher Frikha, Rte MZL Chaker km 1.8, Rue des Roses, SFAX, 3072, Tunisia Tel: 0021624052530

Therefore, the main goal of this research is to reduce the mismatch between training and testing speech by some form of enhancement techniques and consequently to improve the recognition performance. We focused on two kinds of speech enhancement techniques which attempt to suppress the noise from the testing speech. The first one, named spectral subtraction, is not recent approach since it was first introduced in late 70's. But it is still popular since most single microphone noise reduction algorithms in the last decades are based on this technique, which has become almost standard in noise reduction (Malca *et al*., 1996; Okazaki *et al*., 2004). The second is different from the previous one as far as it concerns the application of the discrete wavelet transform in the front end processing stage of the recognition system. The motivation of using such technique comes from the fact that number of recent theoretical studies have found that the orthogonal wavelet transform offers a promising approach to noise removal (Farooq and Datta, 2003; Yi and Loizoru, 2004).

SPEECH ENHANCEMENT TECHNIQUES

A typical speech enhancement scheme introduced in a recognition system is shown in Fig. 1.

Where in the recognition process, speech enhancement techniques tend to suppress the noise which corrupts the speech signal before it is fed to the recognizer. All these techniques are generally designed to recover the clean speech signal by improving its Signal-to-Noise Ratio (SNR) which is defined as:

$$SNR = 10 \log_{10} \frac{P_s}{P_N} \quad (1)$$

Where:

P_S = The power spectrum of the speech signal
P_N = The power spectrum of the noise

The next subsections, formally introduce the two speech enhancement techniques adopted in this present study.

Spectral subtraction technique: When the noise process is stationary and speech activity can be detected, Spectral Subtraction (SS) is a direct way to enhance the noisy speech. The block diagram describing the spectral subtraction overall process is shown in Fig. 2,
Where:

- A noisy signal is overlap partitioned in short time of milliseconds which are transformed to the frequency domain by a Fast Fourier Transform (FFT).
- An estimated magnitude spectrum which is usually updated in speech frames is subtracted from each noisy magnitude.
- The noise reduced spectra are transformed back to the time domain using the unchanged phase of the noisy signal and overlap added to give the noise output signal.

From the theoretical view point, we assume that we have a speech signal s(k) corrupted by an additive noise n(k). Noise is supposed to be uncorrelated with the speech and to be non stationary. Therefore, in this case, it is possible to write:

$$d(k) = s(k) + n(k) \quad (2)$$

The amplitude spectrum $X(\omega)$ (in the frequency domain) of a speech signal x(k) can be derived by taking the Discrete Fourier Transform (DFT) of x(k). Its power spectrum is then derived as:

$$P_x(\omega) = |X(\omega)|^2 \quad (3)$$

Since the noise power spectrum $P_N(\omega)$ cannot be directly obtained, a noise power spectrum estimate is calculated by taking its average value during non-speech activity period (Gökhun and Ozer, 2000):

$$\hat{P}_x(\omega) = E\{P_N(\omega)\} \quad (4)$$

The hat symbol on letter stands for the estimation of a signal or spectrum the letter represents.

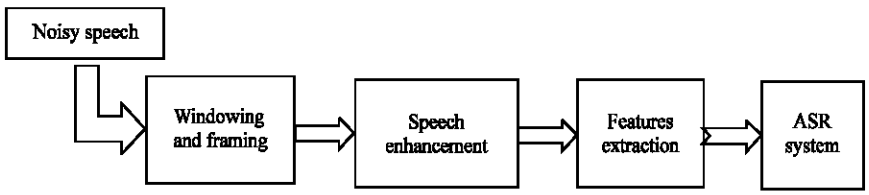

Fig. 1: Speech enhancement scheme for the extraction of robust features

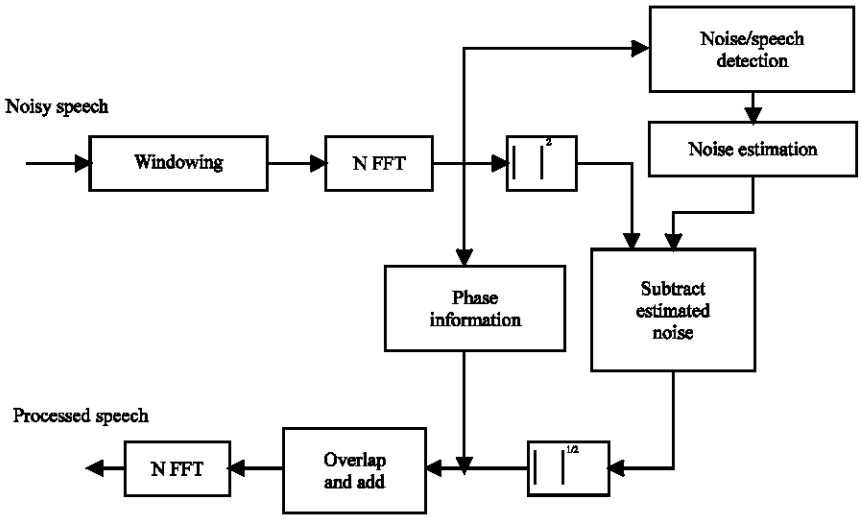

Fig. 2: Spectral subtraction enhancement system bloc diagram

Since the noise is uncorrelated with the speech signal, an estimate of the modified speech spectrum can be obtained by subtracting the noise power spectrum estimate from the corrupt power spectrum:

$$\hat{P}_S(\omega)=P_D(\omega)-\hat{P}_N(\omega) \quad (5)$$

From Eq. 5, it can be seen that the subtraction process involves the subtraction of an averaged estimate of the noise power spectrum from the noisy power speech spectrum. Due to the error in computing the noise power spectrum, we may have some negative values in the modified spectrum. This can cause the estimated of the clean signal to contain musical tones that can be annoying to the listener. This problem can be solved by means of half wave rectification. With half-wave rectification the modified spectrum can be written as:

$$\hat{P}_S(\omega)=\begin{cases}P_S(\omega) \text{ if } \hat{P}_S(\omega)>0\\ 0 \qquad \text{else}\end{cases} \quad (6)$$

The phase spectrum $\varphi_D(\omega)$ calculated from the noisy speech signal is used for reconstruction of the estimated signal spectrum based on the fact that for human perception the short time spectral amplitude is more important than the phase for intelligibility and quality. This conclusion was made by Wang and Lim (1982) in their study, when using the actual phase rather than the degraded speech phase does not improve the quality of the enhanced speech. Since the phase spectrum $\varphi_D(\omega)$ is retained ($\varphi_D(\omega) = \hat{\varphi}_s(\omega)$), the estimated complex spectrum magnitude of the clean speech signal can be calculated:

$$\hat{S}(\omega)=\sqrt{P_D(\omega)-\hat{P}_N(\omega)}\ \exp(\varphi_D(\omega)) \quad (7)$$

The time domain reconstruction of the clean speech signal is then resynthesised through the use of an Inverse Discrete Fourier Transform (IDFT) in conjunction with the overlap and add (OLA) method:

$$\hat{s}(k)=\text{IDFT}(\hat{S}(\omega)) \quad (8)$$

Nearly all later works have found that improved results are obtained by employing noise over-estimates and noise floors (Udrea *et al*., 2005). These ideas were first introduced by the early original work of Berouti (Berouti *et al*., 1979). Equation 5 is thus transformed as follows:

$$\hat{P}_S(\omega)=P_D(\omega)-\alpha\hat{P}_N(\omega) \quad (9)$$

$$\hat{P}_S(\omega)=\begin{cases}P_D(\omega), & \text{if } P_D(\omega)>\hat{P}_N(\omega)\\ \beta\hat{P}_N(\omega), & \text{otherwise}\end{cases} \quad (10)$$

Where, $\alpha>1$ minimizes the appearance of negative values that generate spectral spikes and $0<\beta<<1$ sets a spectral flooring which reduces the perception of musical noise. The optimal value for α can be set as a function of the SNR, as high SNR frames need less compensation than low SNR frames.

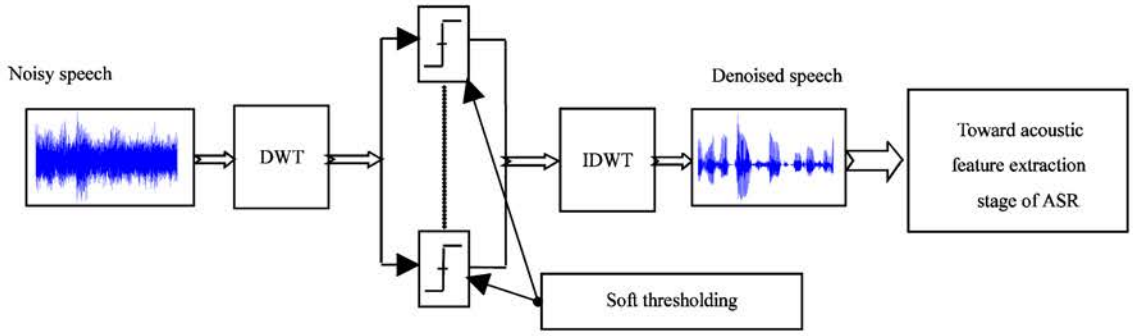

Fig. 3: Bloc diagram of the enhancement preprocessing stage based on DWT

Discrete Wavelet denoising technique: In Fourier based signal processing; the out of band noise can be removed by applying a linear time invariant filtering approach. However, it cannot be removed from the portions where it overlaps the signal spectrum. The denoising technique used in the wavelet analysis is based on an entirely different idea and assumes the amplitude rather than the location of the spectrum of the signal to be different from the noise. The localising property of the wavelet is helpful in thresholding and shrinking the wavelet coefficients that helps in separating the signal from noise. Donoho and Johnston (1995) were first to formalize the wavelet coefficient thresholding for removal of additive noise from deterministic signal. The denoising by wavelet is quite different from traditional filtering approaches because it is non linear, due to a thresholding step. Figure 3 shows the block diagram of the denoising process based on thresholding the discrete wavelet coefficients.

The denoising by thresholding of a signal d(k) contaminated by additive noise (Eq. 2) is performed as follows:

- Perform the wavelet transform of the noisy data.
- Calculate the threshold δ depending upon the noise variance.
- Perform thresholding of the wavelet coefficients.
- The coefficients obtained from step 3 are then padded with zeros to produce a legitimate wavelet transform and this is inverted to obtain the signal estimate.

The threshold δ is calculated using the signal obtained from the high pass filter output (detailed coefficients) according to Eq. 11:

$$\delta = s\sqrt{2.\log(n)} \quad (11)$$

Where:

n = The size of the data used to calculate the threshold

s = The estimation of the noise done by using median absolute deviation (Donoho and Johnston, 1995; Yi and Loizou, 2004)

Usually thresholding is applied on the detailed coefficients and the approximate coefficients (the low pass filter output) are left untouched.

Mathematically, for the detailed coefficient d_{ij}, the thresholding is carried out as follows:

$$\hat{d}_{ij} = \begin{cases} \text{sign}(d_{ij}).(|d_{ij}| - \delta) & \text{if } |d_{ij}| > \delta \\ 0 & \text{if } |d_{ij}| \leq \delta \end{cases} \quad (12)$$

Where, sign(x) is 1 if x is positive and -1 if x is negative. The technique of soft thresholding is also called wavelet shrinkage because all the wavelet coefficients are reduced. Shrinkage of the wavelet coefficients is more helpful in reducing the noise from the signal as compared to the hard thresholding method. The extent of denoising depends upon the level of decomposition. For higher level of decomposition, denoising can be applied to all the detailed coefficients. It is possible that some of the signal information may also be lost during the denoising process and the loss increases with the increase in the level of decomposition. The mother wavelet chosen for denoising was Daubechies 4. The signal after denoising is smoother which also causes the removal of some of the signal components. This may cause reduction in the recognition performance at higher signal to noise ratios for the phonemes having high frequency components (e.g., fricatives).

RESULTS

Speech recognition conditions: A small vocabulary isolated word task based on TIMIT database (Darpa,

1990) was used for evaluation. This recognizer has been studied and optimized in our previous research work, (Frikha *et al.*, 2007). The TIMIT database was recorded in quiet conditions and sampled at 16 KHz. Model training and evaluation were performed using HTK 3.2 program package (Young *et al.*, 2002). The performance of the recogniser was measured in terms of word accuracy (WAC):

$$WAC = \frac{N - S - D - I}{N} \times 100\% \quad (13)$$

Where:

N = The total No. of words in the test set
S = The No. of substitution errors
D = The No. of deletion errors
I = The No. of insertion errors

Ten words selected from the two sentences sa1 and sa2 from the eight dialect region (DR1… DR8) were used for the training and testing the recognizer. Each vocabulary word was modelled by 3 emitting states left to right HMM with one Gaussian component per state and no skip transition. The training set contains 4620 words whereas the testing set contains 1680 words (respectively 462 and 168 repetitions of each of the 10 words in the vocabulary).

For the purpose of evaluating the robustness against environmental additive noise, 4 different types of noise were added to the test data with different Signal to Noise Ratios (SNR) including 0, 5, 10 and 20 dB, while those of the training data are kept free of noise. The typical types of noise include white Gaussian noise, babble noise, factory noise and pink noise all extracted from NOISEX database (Varga *et al.*, 1992). Noisy speech was generated in the following way: for each speech file in the test corpus, a noise segment of length equal to the length of the speech file was randomly extracted, multiplied by a gain factor which depends on the desired SNR and added to the speech file.

Four basic kinds of feature vector in the acoustic front-end of the isolated word recognition system were considered (MFCC, PLP, LPC and LPCC). Those vectors were computed, using the waveform analysis tools provided with HTK, every 10 ms using 25 ms Hamming analysis window.

The acoustic model for a given word is chosen to be Hidden Markovian (HMM) (Rabiner, 1989). The estimation of the parameter sets of the HMMs is usually performed using the Expectation-Maximization (EM) algorithm by the Maximum Likelihood (ML) function of the HMM (Dempster *et al.*, 1977) for a given sequence of speech signal.

Experimental results: Comparative experiments are conducted with the previously mentioned acoustic front end features. Typically, a speech recognition feature vector consists of 12 static coefficients (C_1, C_2,…, C_{12}) to which might be added one of the following component: a log energy (_E), first derivative (_D), log energy and first derivative (_D_E), first derivative and second derivative (_D_A), log energy, first derivative and second derivative (_E_D_A). It is believed that the addition of the first and second derivatives of the static features should ameliorate the performance of the recognizer (Furui, 1986).

The goal of our first experiment was to study the performance of each kind of feature in clean environments (Table 1).

As can be noticed and for all kinds of features, the best performance of the isolated word recognition system is obtained with static parameters appended by log energy of the frame and their first derivative components, since for that kind of feature, we get an overall relative improvement in performance of 9.5% over static features. Therefore, we adopt this kind of features for our remaining experiments. Also, we noticed the poor performance of the LPC features in comparison with MFCC, PLP and LPCC. The second experiment was targeted on the study of the performance of the recognition system in noisy environments. The four kinds of acoustic features augmented by the log energy of the frame and first derivative components were maintained (Table 2).

From the obtained results, we noticed the degradation of the performance of the recognition system caused by the mismatch between training and testing conditions especially at low SNR levels. We believe that this is due to the fact that, at such SNR levels, it is difficult to estimate accurately clean features from noisy speech because the Expectation-Maximization (EM) algorithm may converge to a wrong solution if the mismatch between training and testing conditions is too large.

The large mismatch causes a bad initial condition of EM algorithm and in turn leads the EM algorithm to converge to unexpected point (Ephraim and Merhav, 2002).

Table 1: Performance of the recognition system for different acoustic features

Acoustic features	Static	E	D	E D	D A	E D A
MFCC	98.21	98.69	98.87	98.9	98.87	98.63
PLP	98.27	98.57	98.87	99.0	98.87	98.75
LPCC	96.96	97.86	98.69	98.7	98.57	98.39
LPC	81.60	85.53	26.19	87.3	79.69	85.71

Table 2: Performance of the system tested under four noise conditions

	Feature kind			
	MFCC_E_D	PLP_E_D	LPC_E_D	LPCC_E_D
	Environment Word Accuracy (WAC) in (%)			
	Clean			
Noise conditions	98.93	98.99	87.25	98.69
SNR = 20 dB				
Pink	85.71	85.17	43.72	81.77
Factory	87.02	86.72	45.03	86.24
Babble	88.33	88.27	62.42	87.97
White	77.13	77.13	48.42	70.10
Average	84.55	84.32	49.90	81.52
SNR = 10 dB				
Pink	57.73	57.18	26.98	50.80
Factory	71.05	70.52	30.08	66.89
Babble	86.72	87.14	40.08	86.42
White	46.75	46.75	29.12	33.00
Average	65.56	65.40	31.57	59.28
SNR = 5 dB				
Pink	39.31	37.76	18.58	30.67
Factory	48.18	47.65	19.00	41.87
Babble	79.99	82.61	22.99	73.14
White	34.04	34.07	22.87	24.78
Average	50.38	50.52	20.86	42.62
SNR = 0 dB				
Pink	31.51	27.58	12.33	24.06
Factory	34.66	32.70	14.71	29.24
Babble	57.30	58.49	13.04	39.19
White	28.65	28.65	17.57	22.69
Average	38.03	36.86	14.41	28.80

Table 3: Performance of the enhancement techniques based on Spectral Subtraction (SS) and Discrete Wavelet Transform (DWT)

		Word accuracy (%)		
		98.93		
		Clean	Enhancement technique	
Environments		Noisy	SS	DWT
Average	20 dB	84.55	93.67	95.83
	10 dB	65.56	79.48	80.11
	5 dB	50.38	66.50	64.85
	0 dB	38.03	51.29	47.04

The goal of our final experiment was basically to investigate the two enhancement techniques theoretically. Those techniques are based on the Spectral Subtraction (SS) justified by Berouti *et al.* (1979) study and a novel discrete wavelet (DWT) procedure based on Farooq and Datta (2001) study. The algorithms inherent to those techniques were all implemented in Matlab. We've been interested only on the feature kind MFCC_E_D since it leads to the best performance according to our previous experiments (Table 3).

EVALUATION RESULTS AND DISCUSSION

Our first experiment is conducted to evaluate the effectiveness of several signal processing schemes used as acoustic front ends of an isolated word recognition system in clean environments. Four parametric representations of acoustic signal were compared: MFCC, PLP, LPCC and LPC. Those static parameters were eventually appended by the frame's log energy and their first and second derivatives coefficients. Results showed that best word accuracy obtained for static features augmented by the frame's log energy and their first derivatives. We also noticed the poor performance of LPC front end when compared with the cepstrum and the perceptual parameters (MFCC, PLP and LPCC). It is believed that those kinds of parameters succeed better than LPC in capturing the relevant information to the recognition system (Jankowski *et al.*, 1995). Also, it is worth noting the almost same recognition performance obtained by the perceptually based acoustic features (MFCC and PLP).

Our second experiment is intended to compare the robustness of the four static acoustic front ends appended by the frame's log energy and their first derivatives coefficients. We aimed to provide information of those features to noise. The HMM isolated word recognition system was therefore tested under four additive noise conditions. The average relative loss of performance in % of the recognition system over to that performed in clean conditions is summarized in Table 4.

From this experiment, we point out the significant average loss of performance of the recognition system which obviously depends on the SNR level. However, when the perceptual appended acoustic features (MFCC_E_D and PLP_E_D) are considered, the observed average loss of performance is within 15% to 65% for SNR ranging from 20 to 0 dB. Moreover, LPC based front end seems to be not immune to noise since with such kind of feature, the average relative loss of performance significantly drops to 43% for SNR= 20 dB and reaching 84% for SNR=0 dB.

Finally, two pre-processing speech enhancement techniques respectively based on Spectral Subtraction (SS) and Discrete Wavelet Transform (DWT) were evaluated. Only the MFCC_E_D acoustic features were considered for this experiment. The average relative recognition rate improvement of the implemented enhancement algorithms with regard to the performance of the recognition system in noisy environment is shown in Table 5.

From the obtained results, it is worthwhile to note the best performance of the SS over the DWT enhancement technique at low SNR levels (0 and 5 dB). In fact, the SS technique assures an average improvement rate of performance around 35% and around 32%, respectively

Table 4: Relative loss of performance of the recognition system in noisy environments

	Average relative loss of performance (%)			
	Feature kind			
SNR	MFCC_E_D	PLP_E_D	LPCC_E_D	LPC_E_D
20 dB	14.5	15.0	17.5	42.8
10 dB	34.0	34.0	40.0	64.0
5 dB	49.1	49.0	57.0	76.1
0 dB	61.6	65.6	70.1	83.5

Table 5: Recognition rate improvement brought by the implemented speech enhancement techniques

	Recognition rate improvement (%)	
	Enhancement technique	
SNR	SS	DWT
20 dB	10.8	13.4
10 dB	21.3	22 .2
5 dB	32.0	28.8
0 dB	34.9	23.7

for SNR of 0 and 5 dB. At high SNR levels (10 and 20 dB), the improvement brought by the DWT enhancement technique is slightly better than that of the SS. We believe that the signal, obtained after denoising with the DWT enhancement technique, is smoother which causes the removal of some of the signal components at high noisy conditions. This may cause reduction in the recognition performance at lower signal to noise ratios for the phonemes having high frequency components (e.g., fricatives).

CONCLUSIONS

In this study, we studied two categories of robust speech recognition problem from the view point of parametric representations of speech characteristics and front end acoustic stage compensation. Four standard static acoustic features (MFCC, PLP, LPCC and LPC) eventually appended by log energy of the frame and their first and second derivative components and two speech enhancement techniques based on spectral subtraction and discrete wavelet transform were studied in a speaker independent isolated word recognition tasks. Several experiments have been carried out in both clean and additive noisy environments. Following are some of our findings:

- Best performance of the recognition system is obtained with static parameters appended by log energy of the frame and their first derivative components, since for that kind of feature; we get an overall relative improvement of performance of 9.5% over static features.
- Almost the same recognition results are obtained with the perceptually based acoustic features (MFCC and PLP). Moreover, those representations outperform the LPC and LPCC characteristics.
- When tested under additive noisy conditions, an average loss of performance of the recognition system of about 15 to 65% for SNR ranging from 20 to 0 dB is observed when the perceptual appended acoustic features (MFCC_E_D and PLP_E_D) are used. Moreover, LPC features seems to be not immune to noise since with such parameters, the word accuracy of the recognizer significantly drops to 43% for SNR = 20 and to 84% for SNR = 0 dB.
- It is shown that the utilisation of the speech enhancement technique based on discrete wavelet transform does not result in a clear advantage over spectral subtraction when used in our application at low SNR values (0 and 5 dB). At high SNR levels (10 and 20 dB), the discrete wavelet transform enhancement technique achieves slightly better recognition rate improvement than that of spectral subtraction.

REFERENCES

Beroutti, M., R. Schwartz and J. Makhoul, 1979. Enhancement of speech corrupted by acoustic noise. In: Proceeding International Conference Acoustics, Speech, Signal Processing, 1: 208-211.

Choi, E.H.C., 2004. Noise Robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. Proceeding of the 10th Australian International Conference on Speech Science and Technology, Macquarie University, Sydney, pp: 451-456.

Cui, X. and A. Alwan, 2005. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR. IEEE Trans. Speech Audio Process, 13 (6): 1161-1172.

Darpa, 1990. Timit acoustic-phonetic continuous speech corpus (TIMIT) Training and Test Data and Speech Header Software NIST Speech Disc CD1-1.1 October.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., 39: 1-38.

Donoho, D.L. and I.M. Johnston, 1995. De-noising by soft-thresholding. IEEE Trans. Inform. Theory, 41 (3): 613-627.

Ephraim, Y.E. and N. Merhav, 2002. Hidden Markov processes. IEEE Trans. Inform. Theory, 48 (6): 1518-1569.

Farooq, O. and S. Datta, 2001. Robust features for speech recognition based on admissible wavelet packet. IEEE Electron. Lett., 37 (5): 1554-1556.

Farooq, O. and S. Datta, 2003. Wavelet-based denoising for robust feature extraction for speech recognition. IEEE Electron. Lett., 39 (1): 163-165.

Frikha, M., S. Ben Massaoud, M. Kammoun, D. Gargouri, M. Lahyani and A.B. Hamida, 2005. Optimizing some HMM parameters in an isolated speech recognition system. 3rd IEEE International Conference on Systems, Signals and Devices, Vol. III, CSP.

Frikha, M., Z.B. Massaoud, A. Boubaker and A.B. Hamida, 2007. A study of additive noise removal techniques for robust isolated word recognition. 4th IEEE International Multi-Conference SSD07, Vol III, CSP.

Furui, S., 1986. Speaker independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Proces., 34 (1): 52-59.

Gökhun, S. and H. Özer, 2000. Voice activity detection in nonstationary noise. IEEE Trans. Speech Audio Process, 8 (4): 478-482.

Hegde, R.M., 2005. Fourier transform phase-based features for speech recognition. Ph.D Thesis, Indian Institute of Technology Madras.

Jankowski, C.R., H.D.H. Vo and R.P. Lippmann, 1995. A comparison of signal processing front ends for automatic word recognition. IEEE Trans. Speech Audio Process, 3: 286-293.

Malca, Y., D. Wulich, G. Ramponi, G.L. Sicuranza, S. Carrato and S. Marsi, 1996. Improved spectral subtraction for speech enhancement. Signal Processing VIII, Theories and Applications, Proceeding of Eusipco-96, Trieste, Italy, 2: 975-978.

Okazaki, M., T. Kunimoto and T. Kobayashi, 2004. Multi-stage spectral subtraction for enhancement of audio signals. In: Proceeding Int. Conf. Acoust. Speech Signal Processing, 2: 805-808.

Rabiner, L.R., 1989. A tutorial on Hidden Markov models and selected applications in speech recognition. Proceeding IEEE., 77: 257-286.

Udrea, R.M., S. Ciochina and D.N. Vizireanu, 2005. Reduction of background noise from affected speech using a spectral subtraction algorithm based on masking properties of human ear. Int. Conf. IEEE, Telsiks, 1: 135-138.

Varga, A.P. *et al*., 1992. The NOISEX-92 - Study on the effect of additive noise on an automatic speech recognition. In: Technical Report; DRA Speech Research Unit.

Wang, D.L. and J.S. Lim, 1982. The unimportance of phase in speech enhancement. IEEE Trans. Acoustics Speech and Signal Process, 30 (4): 679-681.

Yi, H. and P.C. Loizou, 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process, 12 (1): 59-67.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Vatchev and P. Woodland, 2002. The HTK Book 3.2, Cambridge University Engineering Department, available: http://www.htk.eng.cam.ac.uk.

Zhang, Z. and S. Furui, 2004. Piecewise-linear transformation-based HMM Adaptation for noisy speech. Speech Commun., 42 (1): 43-58.