14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPECTRAL PEAKS ENHANCEMENT FOR EXTRACTING ROBUST SPEECH FEATURES

Babak Nasersharif[1], Ahmad Akbari[1], Mohammad Mehdi Homayounpour[2]

1: Computer Engineering Department- Iran University of Science and Technology
2: Computer Engineering Department- Amirkabir University of Technology
{nasser_s, akbari}@iust.ac.ir, homayoun@ce.aut.ac.ir

ABSTRACT

It is generally believed that the external noise added to speech signal corrupts speech spectrum and so speech features. This feature corruption degrades speech recognition systems performance. One solution to cope with the speech feature corruption is reducing the noise effects on the speech spectrum. In this paper, we propose to filter speech spectrum in order to enhance its spectral peaks in presence of noise. Then, we extract robust features from the spectrum with enhanced peaks. In addition, we apply the proposed filtering to another form of speech spectral representation known as modified group delay function (GDF). Phoneme and word recognition results show that MFCC features extracted from the spectrum with enhanced peaks are more robust to noise than MFCC derived from main noisy spectrum. In addition, MFCC features extracted from filtered GDF are more robust to noise than other MFCC features, especially in low SNR values.

1. INTRODUCTION

Traditional speech features are typically extracted from power spectrum or amplitude spectrum of speech signal. Then, when speech spectrum is changed due to presence of additive noise, these features show a high sensitivity to the noise. This usually results in performance degradation of speech recognition system in presence of additive noise.

Several techniques have been proposed to reduce sensitivity of features to external noise. In some approaches, a transformation is directly applied to feature vectors to remove noise effects such as, cepstral mean normalization (CMN) [7] and SNR dependent cepstral normalization (SDCN) [7]. Some other methods, work at the spectral level. These methods try to reduce the effect of additive noise on the speech spectrum and then extract features. Spectral subtraction [7] and different spectral filtering techniques are well known examples of such methods. Spectral subtraction, subtracts an estimation of noise spectrum from speech power spectrum to remove noise effects from it. Phase autocorrelation (PAC) is another example of these techniques that is recently introduced. It tries to make autocorrelation coefficient less sensitive to additive noise [4]. In this way, it enhances speech spectral peaks. Group delay function (GDF), negative derivative of speech phase spectrum, is another technique used for speech spectrum estimation [9] and robust feature extraction [3][8]. In group delay function, features derive from speech phase spectrum instead of speech power or amplitude spectrum [3][5].

In this paper, we propose to filter speech spectrum for enhancing its spectral peaks in presence of noise and then extract MFCC features from enhanced spectrum. In this way, we use differential power spectrum (DPS) [6] and PAC spectrum for filtering the speech spectrum. Furthermore, we use a type of modified group delay function as a speech spectral representation less affected by noise. Then, we filter this group delay function by DPS and PAC spectrum to enhance its spectral peaks in presence of noise. In this way, we obtain a spectral representation of speech less affected by noise with enhanced spectral peaks. After that, we derive MFCC features from this obtained spectrum.

The remainder of this paper is organized as follows. Section 2 discusses the differential power spectrum. In section 3, we describe phase autocorrelation spectrum and its properties. In section 4, the group delay function and its modification are explained. In Section 5, we explain our proposed filtering method for enhancing spectral peaks. Section 6 includes our experiments and results. Finally, our conclusions are given in section 7.

2. DIFFERENTIAL POWER SPECTRUM

If we denote the power spectrum of the *i* th frame of speech signal as *Y*(*i*, *k*), the differential power spectrum (DPS) can be defined by following difference equation [6]:

$$D(i,k) = \sum_{l=P1}^{P2} b_l Y(i,k+l) \qquad (1)$$

where P1 and P2 are the orders of differential equations, b_l 's are some real-valued weighting coefficients and $0 \leq k < K$, here K is frame length.

We should to resolve three problems to use DPS in practical speech applications. The first one is the selection of proper orders for difference equations, named as P1 and P2 in equation (1). The second one is the determination of weights b_l's in (1). The third one is how DPS should be converted to into parameters that can represent information included in a speech signal which is necessary for recognition purpose.

Unfortunately, an optimal solution to any of the three listed problems is difficult to obtain. We will show here only empirical solutions for practical application [6]. For the first two problems, we can define three special forms of DPS as following:

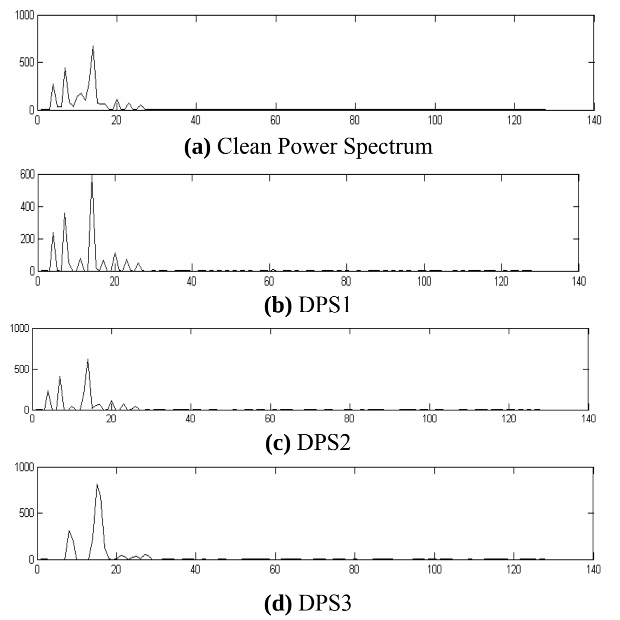
$$D(i,k) = Y(k) - Y(k+1) \qquad (2)$$

$$D(i,k) = Y(k) - Y(k+2) \qquad (3)$$

$$D(i,k) = Y(k-2) + Y(k-1) - Y(k+1) - Y(k+2) \qquad (4)$$

The third problem has been solved in [2][6] by converting DPS into cepstral coefficients. For this purpose, an absolute operation is applied to DPS to make its negative parts positive. Then, MFCC features are extracted from DPS magnitude. Based on results in [6], MFCC features extracted from DPS give the higher recognition rate than MFCC extracted from power spectrum on TI46 database.

Fig. 1 shows power spectrum and different differential power spectra defined in equations (2), (3) and (4) for a sample speech frame corresponding to phoneme /ow/. In the figure, the DPS magnitude is shown after removing its negative values. As shown in the figure, DPS1 represents spectral peaks more accurately than DPS2 and DPS3.



(a) Clean Power Spectrum

(b) DPS1

(c) DPS2

(d) DPS3

Fig.1. Power spectrum and different differential spectra for a sample frame of phoneme /ow/ where sampling frequency is 16 kHz**.**

In this paper, we apply DPS to construct a filter for speech power spectrum in order to enhance peaks of power spectrum in presence of noise. Based on results in [6] and our observations, we have chosen DPS1 to construct our filter.

3. PHASE AUTOCORRELATION

Traditional autocorrelation function is computed as a dot product between the time delayed speech vectors. Recently, an alternative measure of autocorrelation called phase autocorrelation (PAC) has been introduced, where the angle between the vectors in the signal vector space is used as a measure of autocorrelation [4]. The motivation for the use of angle is the fact that angle gets less affected in the presence of noise than the dot product [11].

We give a short review of the Phase AutoCorrelation (PAC), firstly presented in [3], in the following. Consider a speech frame s as:

$$s = \{s[0], s[1], .. s[N-1]\}\} \qquad (5)$$

where N is the frame length. Suppose two vectors x_0 and x_k as:

$$\begin{aligned} x_0 &= \{s[0], s[1], .. s[N-1]\}\} \\ x_k &= \{s[k], ..., s[N-1], s[0], ..., s[k-1]\} \end{aligned} \qquad (6)$$

Then, the autocorrelation coefficients of the speech frame are computed using dot product by:

$$R[k] = x_0^T x_k \qquad (7)$$

On the other hand, R[k] can be shown by:

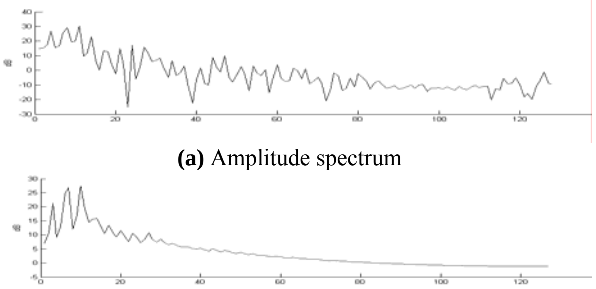
$$R[k] = |x|^2 \cos(\theta_k) \qquad (8)$$

where $|x|^2$ denotes the energy of the frame and Θ_k represents the angle between vectors x_0 and x_k in N dimensional space. PAC coefficients are derived from autocorrelation coefficients as below:

$$P[k] = \theta_k = Arc \cos \left(\frac{R[k]}{|x|^2}\right) \qquad (9)$$

As angle gets less affected in noise than dot product, PAC coefficients are more robust to noise than the regular autocorrelation coefficients [11]. The Fourier equivalent of PAC coefficients in frequency domain is called PAC spectrum. The computation of PAC coefficients from the autocorrelation coefficients using (9) includes 2 operations: energy normalization and inverse cosine. As explained in [4], the inverse cosine transformation has an effect of enhancing the spectral peaks out of spectral valleys. This can be seen in the Figs. 2(a) and 2(b), where regular and the PAC spectrum are shown for a sample speech frame corresponding to phoneme /ow/. According to Fig. 1, PAC in one hand enhances spectral peaks, and in other hand it gives less weight to some high frequency information of the spectrum. Due to this, PAC spectrum does not include the same information of clean spectrum, but it can enhance the noisy spectrum and, especially, its peaks. Therefore, if we apply PAC spectrum as a filter to power spectrum, it can enhance the power spectrum spectral peak and save its details. This is shown in the Fig. 4(e).

Similar to the features extracted from the regular spectrum, a class of features can be extracted from the PAC spectrum. Mel frequency cepstral coefficients extracted from PAC spectrum is called PAC-MFCC. Experimental results in [1] and [4] show that PAC-MFCC is very robust to noise but it does not work well in clean speech conditions.



(a) Amplitude spectrum

(b) PAC spectrum

Fig. 2. Amplitude spectrum and PAC spectrum for a sample frame of phoneme /ow/**.** Sampling frequency is 16 kHz.

4. GROUP DELAY FUNCTION

It is widely perceived that the magnitude spectrum visually represents the speech spectral information much better than phase spectrum. It is interesting that unlike the phase spectrum, its negative derivative, called the group delay function (GDF) [3][5] [9], can be effectively used to extract various speech signal parameters when the signal under consideration is a minimum phase signal. This is due to fact that the magnitude spectrum of a minimum phase signal and its group delay function are similar to each other. The group delay function is defined as:

$$\tau_p(\omega) = -\frac{d(\theta(\omega))}{d\omega} \qquad (10)$$

where θ (ω) is the unwrapped phase function. The group delay function can also be calculated from the speech signal, by:

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \qquad (11)$$

where the subscripts R and I indicate the real and imaginary parts, respectively and X(ω) and Y(ω) are the Fourier transforms of x(n) and nx(n), respectively. The group delay function requires that the signal be minimum phase or that the poles of the transfer function be within the unit circle. The group delay function becomes spiky in nature due to pitch peaks, noise and window effects. This has been illustrated in [3] and [5]. It is also noticeable that the denominator in equation (11) vanishes, at zeros that are located close to the unit circle. The next task is therefore to suppress the zeros. The spiky nature of the group delay spectrum can be overcome by replacing the denominator of the group delay function with its cepstrally smoothed version S(ω). This gives the modified group delay function (MGDF) as follows [3][5]:

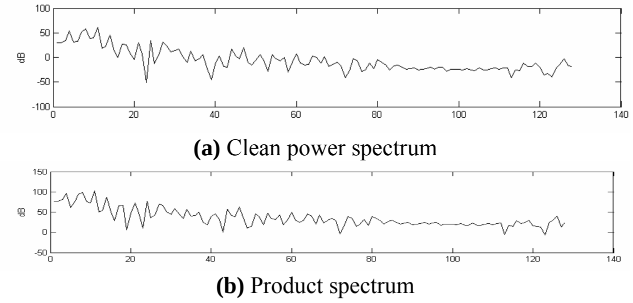
$$\tilde{\tau}_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{(S(\omega))^2} \qquad (12)$$

In [3], Zhu and Paliwal defined the ***product spectrum*** as the product of power spectrum and group delay function as follows:

$$Q(\omega) = |X(\omega)|^2\, \tau_p(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \qquad (13)$$

The product spectrum, called as ***PG*** in this work, is affected by both the magnitude spectrum and the phase spectrum. It enhances the region at the formants over the MGDF and has an envelope comparable to that of the power spectrum. Figs. 3(a) and 3(b) show clean power spectrum and the product spectrum for a sample speech frame corresponding to phoneme /ow/. As shown in Fig. 3(b), the product spectrum represents well the details of clean speech power spectrum, but it can not enhance spectral peaks as well as PAC.

It is shown in [3] that by using MFCC extracted from product spectrum, (named here as ***PG-MFCC***), a higher recognition rate can be obtained than using MFCC extracted from modified group delay function. So, in this work we used PG-MFCC as our recognition features.



(a) Clean power spectrum

(b) Product spectrum

Fig. 3 clean power spectrum and product spectrum for a sample frame of phoneme /ow/ where sampling frequency is 16 kHz.

5. SPECTRAL PEAKS ENHANCEMENT

As mentioned in [6], spectral peaks convey the most important information in speech signal. In addition, they are affected by noise less than other parts of speech signal. Due to this, we believe that amplification of spectral peaks (and then increasing spectral peaks values and peak to valley ratio) can help to reduce noise effects on speech signal and so its features. We showed in section 2 that DPS preserve spectral peaks. Based on this, we use DPS1 as a filter transfer function on power spectrum to enhance the power spectrum peaks in presence of noise. The zero parts of DPS can be considered as 1 in the filter. By this filtering, we reinforce power spectrum spectral peaks and keep its detail. This can be seen in Fig. 4(d) for noisy speech. We name this enhanced power spectrum as ***PDPS***. We also call the MFCC features derived from ***PDPS*** as ***PDPS-MFCC***. We can compute enhanced power spectrum as following equation:

$$PDPS(\omega) = PS(\omega)DPS(\omega)\,|=|\,X(\omega)\,|^2\,DPS(\omega) \qquad (14)$$

where PS(ω), DPS(ω), |X(ω)| are speech power spectrum, differential power spectrum and speech amplitude spectrum, respectively.

It was also showed that PAC spectrum enhances spectral peaks. Therefore, in a similar way, we apply it as a filter transfer function to power spectrum and enhances power spectrum peaks, while preserving its details. This can be seen in Fig. 4(f) for noisy speech. We name this enhanced power spectrum as PPAC. We call the MFCC features extracted from ***PPAC*** as ***PPAC-MFCC***. We can compute enhanced power spectrum as following:

$$PPAC(\omega) = PS(\omega)\,|\,P_a(\omega)\,|=|\,X(\omega)\,|^2\,|\,P_a(\omega)\,| \qquad (15)$$

where PS(ω), |P$_a$(ω)|, |X(ω)| are speech power spectrum, PAC spectrum and speech amplitude spectrum, respectively.

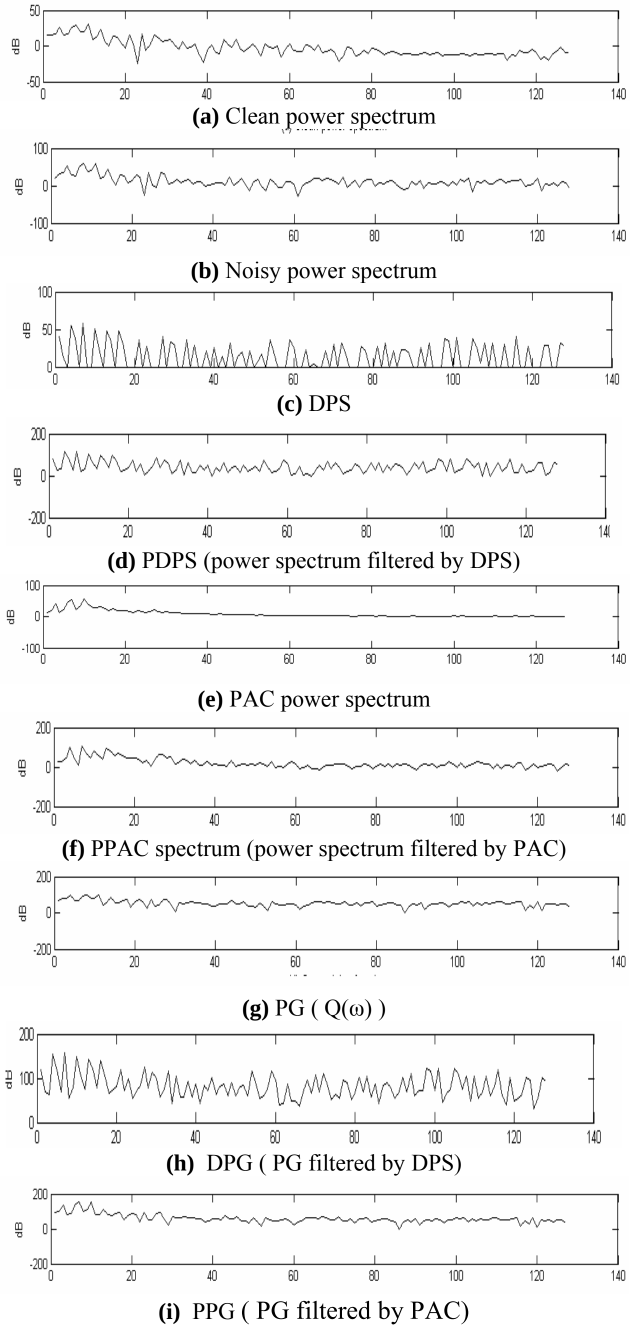
As said in section 4, PG has almost the same information of main speech spectrum. In addition, it is more robust to noise than main speech spectrum as shown in [3]. Consequently, we propose to apply PAC and DPS1 as filters to PG spectrum to enhance its spectral peaks and obtain more robustness to noise. This is shown in Figs. 4(g), 4(h) and 4(i). We named the PG spectrum filtered by PAC and DPS1 as ***PPG*** and ***DPG,*** respectively. We also call the MFCC features derived from PPG and DPG as ***PPG-MFCC*** and ***DPG-MFCC***, respectively. Enhanced spectra can be calculated using following equations:

$$PPG(\omega) = Q(\omega)\,|\,P_a(\omega)\,| = PG(\omega)\,|\,P_a(\omega)\,| \qquad (16)$$

$$DPG(\omega) = Q(\omega)DPS(\omega) = PG(\omega)DPS(\omega) \qquad (17)$$

6. EXPERIMENTS AND RESULTS

We report our results in two parts: phoneme recognition and word recognition. Three types of additive noises were used in both cases: white and pink and factory noises selected from NOISEX92 database. In both of word and phone recognition, our feature vectors contain 12 MFCC and 12 delta-MFCC that are extracted from each of speech main spectrum, PDPS, PAC, PPAC, PG, DPG and PPG spectra. We named these feature vectors as MFCC, DP-MFCC, PAC-MFCC, PPAC-MFCC, PG-MFCC, DPG-MFCC and PPG-MFCC respectively**.**



(a) Clean power spectrum

(b) Noisy power spectrum

(c) DPS

(d) PDPS (power spectrum filtered by DPS)

(e) PAC power spectrum

(f) PPAC spectrum (power spectrum filtered by PAC)

(g) PG (Q(ω))

(h) DPG (PG filtered by DPS)

(i) PPG (PG filtered by PAC)

Fig. 4. Clean power spectrum and different types of noisy power spectra for a sample frame of phoneme /ow/ in presence of white noise with SNR value of 10 dB**.** Sampling frequency is 16 kHz

6.1. Phoneme Recognition

We report our results on TIMIT database for phoneme recognition. We use the same standard train and test set of TIMIT database for phonemes. We divide TIMIT phones to 39 phone classes according to [10]. Sampling frequency of phonemes is 16 kHz. We use CDHMM as recognizer with 3 states and 8 Gaussian mixtures per state which is trained on clean speech**.** We added three mentioned noises to testing set only for phoneme recognition in presence o f noise.

Table 1 shows the average correct rate of phoneme recognition in presence of different noises for different SNR values. According to the table, enhancement of speech spectral peaks using PAC and DPS filters increase the phoneme recognition rate in presence of noise. This can be seen from comparing results of PDPS-MFCC and PPAC-MFCC with MFCC. In addition, PPAC-MFCC has better result than PDPS-MFCC in both of clean and noisy conditions. This shows that PAC is more useful than DPS in filtering of main speech spectrum for enhancing spectral peaks and so robust feature extraction. This is expected, because PAC has more spectral details than DPS. The speech spectral representation based on GDF (PG), improve recognition rate more than other spectra in clean case and SNR Value of 10 dB. But, the best recognition results in SNR value of 0 dB is obtained where we filter PG by PAC spectrum. The filtering of PG using PAC and DPS, increase the phoneme recognition rate in SNR value of 0 dB. In this case, PAC is more useful filter than DPS for enhancing the spectral peaks of PG spectrum.

	Clean	**SNR=10**	**SNR=0**
MFCC	54.20%	38.46%	17.44%
PDPS-MFCC	49.41%	43.49%	19.99%
PAC-MFCC	40.72%	37.27%	21.43%
PPAC-MFCC	**53.18%**	44.10%	20.34%
PG-MFCC	53.49%	46.33%	21.03%
DPG-MFCC	45.45%	42.19%	23.17%
PPG-MFCC	47.66%	**44.80%**	**25.31%**

Table 1. The average of correct phoneme recognition rate in different SNR values for 3 different noise types (factory, pink and white)

6.2. Word Recognition

We also report our results on TIMIT database for isolated word recognition. Two sentences from speakers in two dialect regions were selected and were segmented into words. In this way, we have 21 words spoken by 151 speakers including 49 females and 102 males. These speakers were divided into train and test speakers according to TIMIT speakers division. Our training set contains 2349 utterances spoken by 114 speakers. The testing set includes 777 utterances spoken by 37 speakers. Our recognizer is CDHMM with 6 states and 8 Gaussian mixtures per state which is trained on clean speech.

Fig. 5 and Fig. 6 show average word error rate in presence of 3 mentioned different noises (factory, pink and white) for SNR value**s** of 10 and 0 dB, respectively. The results are reported for all 3216 utterances of testing and training noisy database in terms of word error rate (WER). As shown in figures, applying PAC and DPS filter to main speech spectrum decrease word error rate in both SNR values of 10 and 0 dB. This can be seen from comparing MFCC with PDPS-MFCC and PPAC-MFCC. The best results are obtained when we filter PG by PAC and DPS and so enhance its spectral peaks.

In this case, PAC performs better in filtering. So, PG filtered by

PAC gives the best word recognition results in both SNR values of 10 and 0 dB.

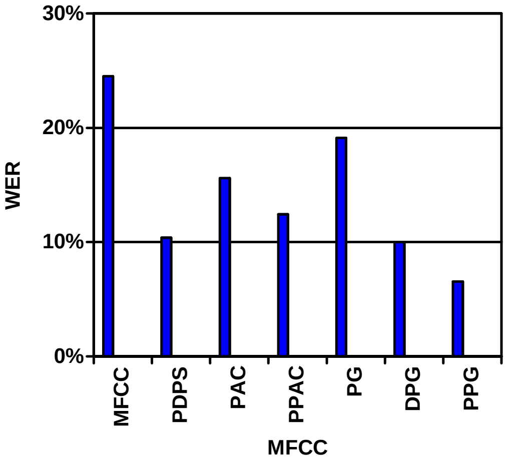


Fig.5. Average word error rates in presence of different noise types for SNR value of **10 dB**

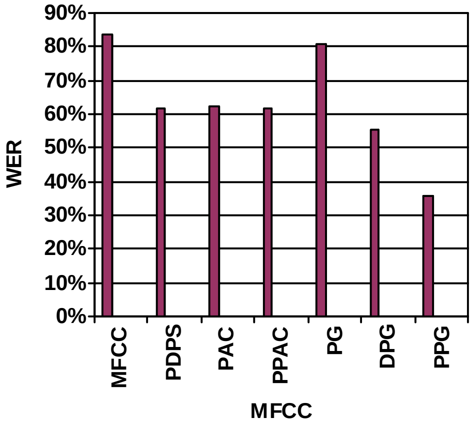


Fig.6. Average word error rates in presence of different noise types for SNR value of **0 dB**

7. CONCLUSION

We proposed to enhance speech spectral peaks in order to obtain more robust speech features. For this purpose, we filtered the speech spectrum by PAC and DPS to enhance its spectral peaks. We used two types of spectra: main speech spectrum and another speech spectral representation based on GDF named here as PG. We filtered both of these spectra by PAC and DPS to enhance their spectral peaks. Phoneme and word recognition results showed that MFCC features extracted from filtered spectra, were more robust to noise than MFCC features derived from the main spectra. Moreover, based on recognition results, PAC performs better than DPS in filtering. In this work, the most robust features and best recognition results has been obtained from PG spectrum filtered by PAC.

As future work, we will try to compare and combine spectral peaks enhancement filters with speech enhancement filters to achieve noise robust speech features.

8. REFERENCES

[1] B.Nasersharif, A.Akbari, "Sub-band weighted projection measure for robust sub-band speech recognition" *Proceeding of EUROSPEECH,* pp. 945-948, 2005.

[2] G. Farahani, S.M. Ahadi, "Robust features for noisy speech recognition based on filtering and spectral peaks in autocorrelation domain" , *Proceeding of EUSIPCO,* 2005.

[3] D. Zhu, K. Paliwal, "Product of power spectrum and group delay function for speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal processing,* vol. 1, pp. 125-128, 2004.

[4] S. Ikbal, H. Misra, H. Bourlard, "Phase autocorrelation derived robust speech features", *IEEE International Conference on Acoustics, Speech, and Signal processing,* vol. 2, pp. 133-136, 2003.

[5] H.A Murthy, V. Gadde, "The modified group delay function and its application to phoneme recognition", *IEEE International Conference on Acoustics, Speech, and Signal processing,* vol. 1, pp. 68-71, 2003.

[6] J.Chen, K.K.Paliwal, S.Nakamura, "Cepstrum derived from differential power spectrum for robust speech recognition", *Speech Communication*, Vol. 41, Issues 2-3, pp. 469-484, October 2003.

[7] X. Huang, A.Acero, H. Hon, *Spoken Language processing*, Prentice Hall, 2001.

[8] A. Bayya, B.Yegnanarayana, "Robust features for speech recognition systems", *Proceeding of ICSLP*, 1998.

[9] B.Yegnanarayana, H.A Murthy, "Significance of group delay functions in spectrum estimation", *IEEE Trans. on Acoustics, Speech and signal processing,* vol. 40, No. 9, pp.2281-2289, September 1992.

[10] K.F. LEE, "Speaker-Independent phone recognition using hidden Markov models", *IEEE Trans. on Acoustics, Speech and signal processing,* vol. 37, pp.1641-1648, 1989.

[11] D. Mansour, B.Juang, "A family of distortion measure based upon projection operation for robust speech recognition", *IEEE Trans. on Acoustic, Speech and signal processing,* vol. 37, pp.1659-1671, 1989.