# ROBUST FEATURES FOR NOISY SPEECH RECOGNITION BASED ON FILTERING AND SPECTRAL PEAKS IN AUTOCORRELATION DOMAIN

*G. Farahani, and S.M. Ahadi*

Electrical Engineering Department, Amirkabir University of Technology
Hafez Avenue, Tehran 15914, Iran
f8023953@aut.ac.ir, sma@aut.ac.ir

## ABSTRACT

This paper introduces a novel representation of speech for the cases where the speech signal is corrupted by additive noises. In this method, the speech features are computed by reducing additive noise effects via an initial filtering stage followed by the extraction of autocorrelation spectrum peaks.

A task of speaker-independent isolated-word recognition was used to demonstrate the efficiency of these robust features. The cases of white noise and colored noise such as factory, babble and car noises were tested. Experimental results show significant improvement in comparison to the results obtained using traditional feature extraction methods.

## 1-INTRODUCTION

Noise robustness is one of the most challenging problems in automatic speech recognition. The performance of ASR systems, trained with clean speech, may drastically degrade in real environments. The main reason for this degradation is the acoustic mismatch between the training and test environments due to environmental effects. The goal of robust speech recognition is to improve the performance of speech recognition in such adverse environments.

The environmental effects are often determined by noise and channel distortion. Noise is additive in spectral domain while channel distortion is multiplicative and therefore appears as additive in logarithmic spectral domain.

An obvious approach to attack the effects of environment is to have a separate training set for each noise type. However, this approach is not practical due to large diversity of noise types encountered in a real environment.

In order to remove the effect of noise, some methods have been found useful and extensively mentioned in the literature such as *SS* (Spectral Subtraction), *NSS* (Non-linear spectral subtraction) [1], *Lin-log RASTA* (Linear-logarithmic RelAtive SpecTrA) [2], *DPS* (Differentiated Power Spectrum) [3], *PMC* (Parallel Model Combination) [4] and *MVDR* (Minimum Variance Distortionless Response) [5].

Furthermore, for suppressing the channel distortion, various techniques have been developed such as *CMN* (Cepstral Mean Normalization), *RASTA* (RelAtive SpecTrAl) [2] and *BE* (Blind Equalization) [6].

Recently, the parameters extracted using an autocorrelation sequence of the noisy signal have been found useful for robust speech recognition. Some examples include magnitude spectrum of higher lag autocorrelation coefficients [7] and RAS (Relative Autocorrelation Sequence) method [8, 9]. Furthermore, according to [10], preserving spectral peaks is very important in obtaining a robust set of features in speech recognition.

We propose a new approach utilizing peaks obtained from autocorrelation spectrum of speech signal. Using such an approach, we found the following results for autocorrelation domain.

1. If our signal is periodic, the autocorrelation will be periodic.
2. Autocorrelation spectrum can well replace signal (power) spectrum for further processing.

Hence, differentiating in autocorrelation spectral domain will preserve autocorrelation spectral peaks, except that each peak is split into two, one positive and one negative. This is similar to what is done in DPS in the spectral domain due to the importance of the spectral peaks. Therefore, we propose the following front-end description for noise robust feature extraction.

Firstly, we calculate the autocorrelation of the noisy signal. If the temporal autocorrelation of noise is a *DC* or slowly varying signal, its effect can be suppressed by a high-pass filter, such as *RAS* filter [8, 9]. Then, following the DPS concept, autocorrelation spectrum is differentiated with respect to frequency. Finally, from the magnitude of differentiated autocorrelation spectrum, we will get a set of cepstral coefficients by passing it through a mel-frequency filter-bank and later to a block of *DCT* (Discrete Cosine Transform). One can expect this method to remove the noise effects in autocorrelation domain by filtering and differentiating the spectrum.

## 2-COEFFICIENTS DERIVED IN AUTOCORRELATION DOMAIN

If *w(t)* is the ambient noise and *x(t)* the noise-free speech signal, then noisy speech signal *y(t)* can be modeled as

$$y(t) = x(t) + w(t) \tag{1}$$

Speech signal is time-variant and non-stationary. Therefore, it is usually analyzed on the discrete domain. Thus we have

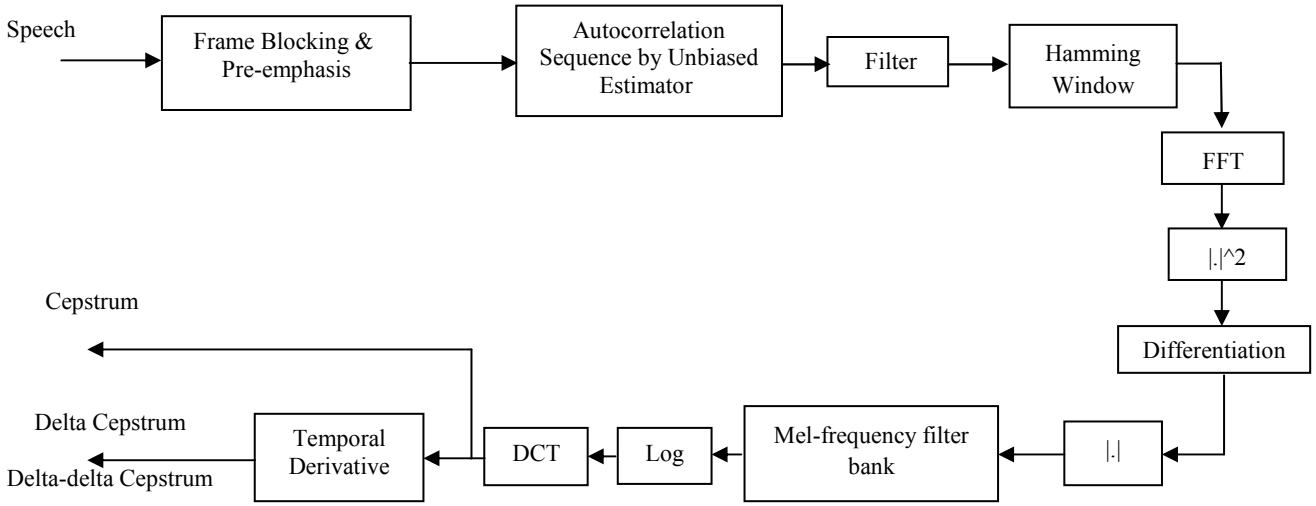


Figure 1. Block diagram of the proposed DAS front-end for robust feature extraction.

$$y(m,n) = x(m,n) + w(m,n)$$
$$0 \le n \le N-1 \quad , \quad 0 \le m \le M-1 \qquad (2)$$

where $N$ is the frame length, $n$ is the discrete time index in a frame, $m$ is the frame index and $M$ is the number of frames.
If the noise is uncorrelated with speech, the autocorrelation of the noisy speech is the sum of the autocorrelation of the clean speech *x(m,n)* and the noise *w(m,n)*.

$$r_y(m,k) = r_x(m,k) + r_w(m,k)$$
$$0 \le m \le M-1 \quad , \quad 0 \le k \le N-1 \qquad (3)$$

where $r_y(m,k)$, $r_x(m,k)$ and $r_w(m,k)$ are the short-time autocorrelation sequences of the noisy speech, clean speech and noise, respectively.
If the noise is stationary, differentiating both sides of (3) with respect to $m$ and simplifying yields [8]:

$$\frac{\partial r_y(m,k)}{\partial m} \cong \frac{1}{T_L}\sum_{t=-L}^{L} t r_y(m+t,k)$$
$$0 \le m \le M-1 \quad , \quad 0 \le k \le N-1 \qquad (4)$$

where

$$T_L = \sum_{t=-L}^{L} t^2$$

Equation (4) is a filtering process on the temporal autocorrelation trajectory by high-pass *FIR* Filter. In our experiments we have chosen $L$=2.
In this step, if the autocorrelation trajectory of noise is *DC* or slowly varying, its effect will be suppressed. Therefore, if we calculate the autocorrelation of signal after this filtering, we will get a cleaner signal, compared to the original noisy signal. After this cleaning step, we can use peaks of autocorrelation signal for feature extraction by differentiating the Fourier transform of autocorrelation of the signal. Thus, if we call the output of the filtering stage in time domain *z(n)*, we can write

$$z(n) = x(n) + v(n) \quad 0 \le n \le N-1 \qquad (5)$$

where *x(n)* and *v(n)* are clean speech and the remaining noise after filtering, respectively. Then, we calculate the autocorrelation function as follows

$$r_z(\tau) = r_x(\tau) + r_v(\tau) \qquad 0 \le \tau \le N-1 \qquad (6)$$

Applying the Fourier transform to both sides of (6) yields

$$F\{r_z(\tau)\} = F\{r_x(\tau)\} + F\{r_v(\tau)\} \qquad (7)$$

or

$$Z(k) = X(k) + V(k) \qquad (8)$$

By differentiating both sides of (8):

$$D_z(k) \approx \sum_{l=-0}^{P} b_l Z(k+l)$$
$$\cong \sum_{l=-P}^{P} b_l \left[X(k+l) + V(k+l)\right] = D_x(k) + D_v(k) \qquad (9)$$

Now differentiation can be carried out in several ways, as discussed in [3]. The simple difference given in (10) has been reported as the best approach and therefore used here.

$$D(k) = Z(k) - Z(k+1) \qquad (10)$$

Our proposed method is different from *RAS* method as we have used the peaks of autocorrelation spectrum found in frequency domain, while it differs DPS in filtering at cleaning step before differentiating the autocorrelation spectrum of the signal. We will call our new features as DAS (Differentiation of Autocorrelation Sequence). The front-end diagram of DAS method is shown in Figure 1.

## 4-EXPERIMENTS

The speech corpus used in these experiments is a speaker-independent isolated-word Farsi (Persian) corpus. The corpus was collected from 65 male and female adult speakers uttering the names of 10 Iranian cities. The data was collected in normal office conditions with SNRs of 25dB or higher and a sampling rate of 16 kHz. Each speaker uttered 5 repetitions of words, some of which were removed from the corpus due to problems that occurred during the recordings. A total of 2665 utterances from 55

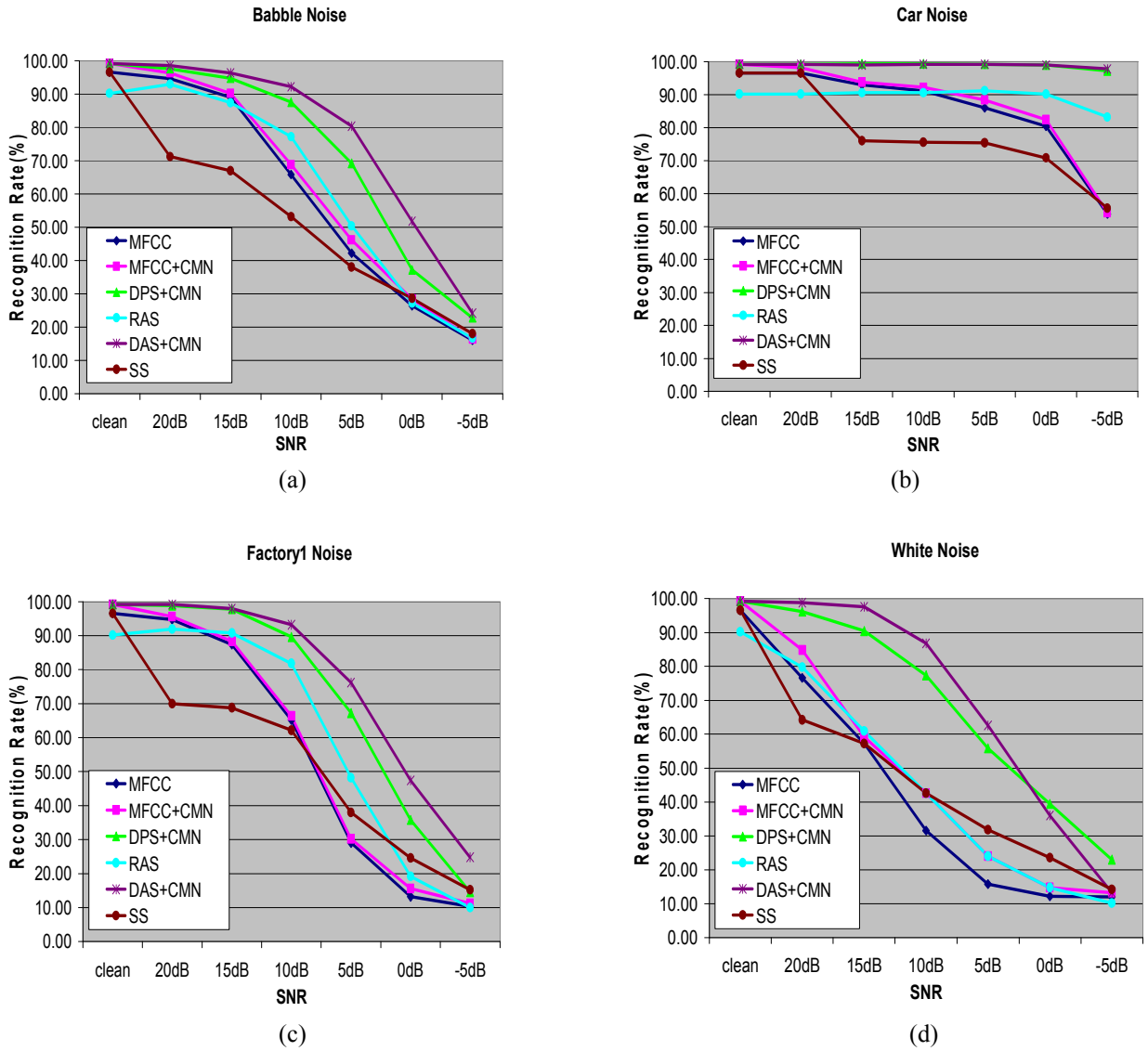


Figure 2. Recognition results for speech signal contaminated with (a) babble, (b) car, (c) factory and (d) white noises in different SNRs. The results correspond to MFCC, MFCC+CMN, DPS+CMN, RAS and DAS+CMN methods.

speakers were used for HMM model training. The test set contained 10 speakers (5 male & 5 female) that were not included in the training set. The noise was then added to the speech in different SNRs. The noise data was extracted from the NATO RSG-10 corpus [11]. We have considered babble, car, factory1 and white noises and added them to the clean signal at 20, 15, 10, 5, 0 and -5 dB SNRs.

Our experiments were carried out using MFCC (for comparison purposes), Spectral Subtraction, RAS-MFCC, DPS and our method (DAS). The features in all cases were computed using 25 msec. frames with 10 msec. of frame shifts. Pre-emphasis coefficient was set to 0.97. For each speech frame, a 24-channel Mel-scale filter-bank was used. Each word was modeled by an 8-state left-right HMM and each state was represented by one Gaussian PDF. The feature vectors were composed of 12 cepstral and a log-energy parameter, together with their first and second derivatives (39 coefficients in total).

Figure 2 depicts the results of our implementation. Also in Table 1 the baseline clean results are included for comparison purposes and in Table 2 the average noisy speech recognition results obtained are displayed. The average values mentioned in Table 2 are calculated over the results obtained from 0 dB to 20 dB SNRs, omitting the clean and -5 dB results. This is the way the average results are calculated in Aurora 2 task [12]. The values given in parentheses are the improvements obtained relative to the baseline system. Note that also for comparison purposes, the results of an implementation of spectral subtraction as an initial enhancement method, applied before standard MFCC parameter extraction, are included. These are denoted by SS and the algorithm was applied as explained in [13]. As can be seen in Figure 2, DAS method outperforms all the other methods in almost all noise types and SNRs. The average results on different SNRs, as shown in Table 2, are again considerably better for DAS in

comparison to other feature extraction techniques. As an example, DAS has about 30% reduction on the average word error rate, compared to DPS, which performs the best among the others. Similar results can be seen on this table for factory and white noises.

## 5-CONCLUSION

In this paper, cepstral features derived from autocorrelation spectral domain were proposed for improving the robustness of ASR systems. The concept of DAS introduced a new set of cepstral features for improving the robustness of speech recognition. We note that just like the RAS and DPS methods, our method can preserve spectral information for speech recognition, while outperforming both RAS and DPS methods due to its two-step noise effect suppression approach. Furthermore, this method works well for different types of noises including white, babble, car and factory noises.

Future works may include the application of an improved filter to the autocorrelation parameters, in place of the current RAS filter.

Table1. Comparison of clean-train/clean-test recognition rates for various feature types .

| Feature type | Recognition Rate |
|---|---|
| MFCC | 96.60 |
| SS | 96.60 |
| MFCC+CMN | 99.20 |
| DPS+CMN | 99.20 |
| RAS-MFCC | 90.20 |
| DAS+CMN | 99.20 |

## ACKNOWLEDGMENT

This work was in part supported by a grant from the Iran Telecommunication Research Center (ITRC).

## REFERENCES

[1] Michael J. Carey, "Robust speech recognition using non-linear spectral smoothing," *Eurospeech'03*, pp. 3045-3048, 2003.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578-589, 1994.

[3] J. Chen, K.K. Paliwal, S. Nakamura, "Cepstrum derived from differentiated power spectrum for robust speech recognition," *Speech Communication*, vol. 41, pp. 469-484, 2003.

[4] M.J.F. Gales and S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comput. Speech Lang.* , vol. 9, pp. 289-307, 1995.

[5] U.H. Yapanel and S.Dharanipragada, "Perceptual MVDR-based cepstral coefficients (PMCCs) for noise robust speech recognition," *ICASSP'03*, pp.644-647, 2003.

[6] L. Mauuary, "Blind equalization in the cepstral domain for robust telephone based speech recognition," *Proc. European Signal Processing Conference*, 1998.

[7] B.J. Shannon and K.K. Paliwal, "MFCC Computation from Magnitude Spectrum of higher lag autocorrelation coefficients for robust speech recognition," *ICSLP2004*.

[8] Kuo-Hwei You and Hsiao-Chuan Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, pp.13-24, 1999.

[9] Kuo-Hwei You and Hsiao-Chuan Wang, "Robust features derived from temporal trajectory filtering for speech recognition under the corruption of additive and convolutional noises," *ICASSP'98*, pp. 577-580, 1998.

[10] B. Strope and A. Alwan, "Robust word recognition using threaded spectral peaks," *Proc. ICASSP'98*, pp. 625-628, 1998.

[11] Available from http://spib.rice.edu/spib/select_noise.html

[12] H-G Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," In *Proc. ISCA ITRW ASR2000,* Paris.

[13] J-C. Junqua and J-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications,* Kluwer Academic Press, Norwell, 1996.

Table2. Comparison of average recognition rates for various feature types with babble, car, factory and white noises.

| Feature type | Average Recognition Rate | | | |
|---|---|---|---|---|
| | Babble | Car | Factory1 | White |
| MFCC (Baseline) | 63.60 | 89.44 | 57.92 | 38.68 |
| SS | 51.60 (-18.87) | 78.88 (-11.81) | 52.72 (-8.98) | 43.88 (13.44) |
| MFCC+CMN | 66.00 (3.78) | 91.00 (1.74) | 59.24 (2.28) | 45.08 (16.55) |
| DPS+CMN | 77.28 (21.51) | 99.24 (10.96) | 77.84 (34.39) | 71.84 (85.73) |
| RAS-MFCC | 67.04 (5.41) | 90.56 (1.25) | 66.40 (14.64) | 44.44 (14.89) |
| DAS+CMN | 83.88 (31.89) | 99.12 (10.82) | 82.80 (42.96) | 76.36 (97.41) |