Feature based on filtering and spectral peaks in autocorrelation domain for robust speech recognition

Author : G.Farahani, S.M.Ahadi, M.M. Homayounpour

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Extraction of robust coefficients in autocorrelation domain
- Experiments

Introduction

- Extra information from different sources usually exists in the speech signal.

  -Noise

  -channel distortion

- One of the main goals of acoustically robust speech recognition is to improve the performance of speech recognition in such adverse environments.

Introduction(cont.)

- we expect this algorithm reduce the noise effects in autocorrelation domain by filtering and differentiating the spectrum and lead to new features that are robust.

-RAS: filtering of the signal autocorrelation is expected to results in the suppression of the effect of noise leading to more robust.

-DPS: differentiation in spectral domain is used to preserve the spectral peaks while the flat parts of the spectrum, that are believed to be believed to be more vulnerable to noise, are almost removed.

Calculation of the autocorrelation domain

- If v(m,n) is the additive noise, x(m,n) noise-free speech signal and impulse response of channel, then the noise speech signal y(m,n) can be written as

$$y(m,n) = x(m,n) + v(m,n),$$

$$0 \le m \le M-1, \quad 0 \le n \le N-1.$$

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 x(m,n), v(m,n) are considered uncorrelated, the autocorrelation can be express as

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(m,k),$$

$$0 \le m \le M-1, \quad 0 \le k \le N-1.$$

where k is autocorrelation sequence index .

Calculation of the autocorrelation domain

- Since addictive noise is assumed to be stationary, its autocorrelation sequence can be considered the same for all frame.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k),$$

$$0 \le m \le M-1,\ 0 \le k \le N-1.$$

- The one-sided autocorrelation sequence of each frame can then be calculate using an unbiased estimator or biased estimator.

$$r_{yy}(m,k) = \frac{1}{N-k} \sum_{j=0}^{N-1-k} y(m,i)\,y(m,i+k)$$

$$r_{yy}(m,k) = \sum_{j=0}^{N-1-k} y(m,i)\,y(m,i+k)$$

Filter of one-side autocorrelation sequence

- After calculating one-sided autocorrelation sequence and differentiating both side with respect to m, the autocorrelation of noise will be deleted and further simplifying yields

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{XX}(m,k)}{\partial m} + \frac{\partial r_{VV}(k)}{\partial m} \cong \frac{\partial r_{XX}(m,k)}{\partial m} = \frac{\sum_{t=-L}^{L} t\, r_{yy}(m+t,k)}{\sum_{t=-L}^{L} t^2},$$

$$0 \le m \le M-1,\ 0 \le k \le N-1,$$

- It is equal to filtering process on the temporal one-sided autocorrelation trajectory by a FIR filter where L is the length of the filter. This filtering process can be written in z domain as follows.

$$H(z) = \frac{\sum_{t=-L}^{L} t z^t}{\sum_{t=-L}^{L} t^2}$$

Calculating differential power spectrum(DPS)

- If the noise and speech signals are assumed mutually uncorrelated, by applying short-time DFT to both side, we can calculate the relation between autocorrelation power spectrums of noise signal and noise as follows:

$$Y(\omega) = FT\{r_{yy}(m,k)\} \approx FT\{r_{xx}(m,k)\} + FT\{r_{vv}(k)\} = X(\omega) + V(\omega)$$

- The differential power spectrum will then be define as

$$Diff_Y(\omega) = \frac{dY(\omega)}{d\omega} = \frac{dX(\omega)}{d\omega} + \frac{dV(\omega)}{d\omega} = Diff_x(\omega) + Diff_v(\omega)$$

Calculating differential power spectrum(DPS)

- In discrete domain, the definition of DPS can be approximated by the following equation

$$Diff_Y(\omega) = Diff_x(\omega) + Diff_v(\omega) \approx \sum_{l=-Q}^{P} a_1 Y(k+l) \cong \sum_{l=-Q}^{P} a_1 [x(k+l) + V(k+l)],$$

$$0 \le k \le K-1$$

where P and Q are the orders of difference equation, a_1 are real-value weighting coefficients and K is the length of FFT.

Description of propose method

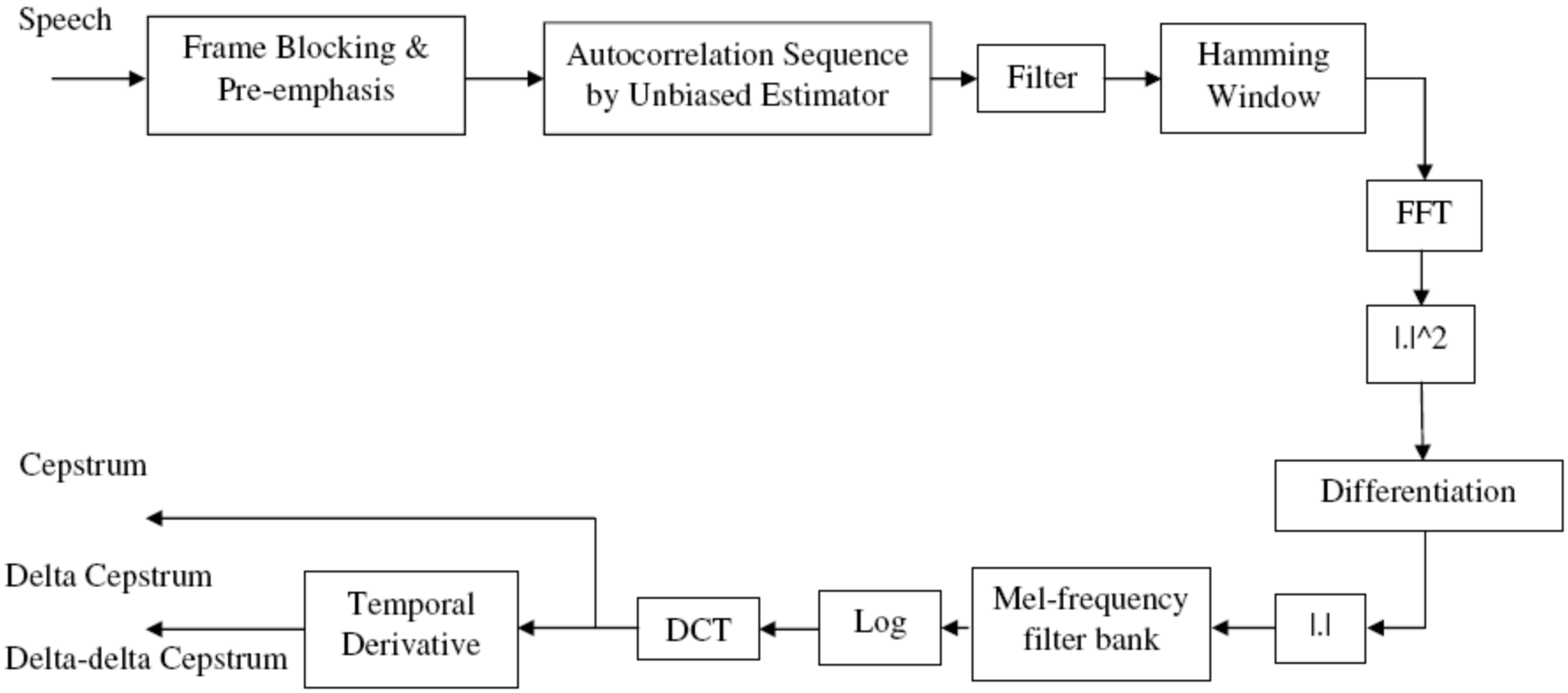


Fig. 2. Block diagram of the proposed DAS front-end for robust feature extraction.

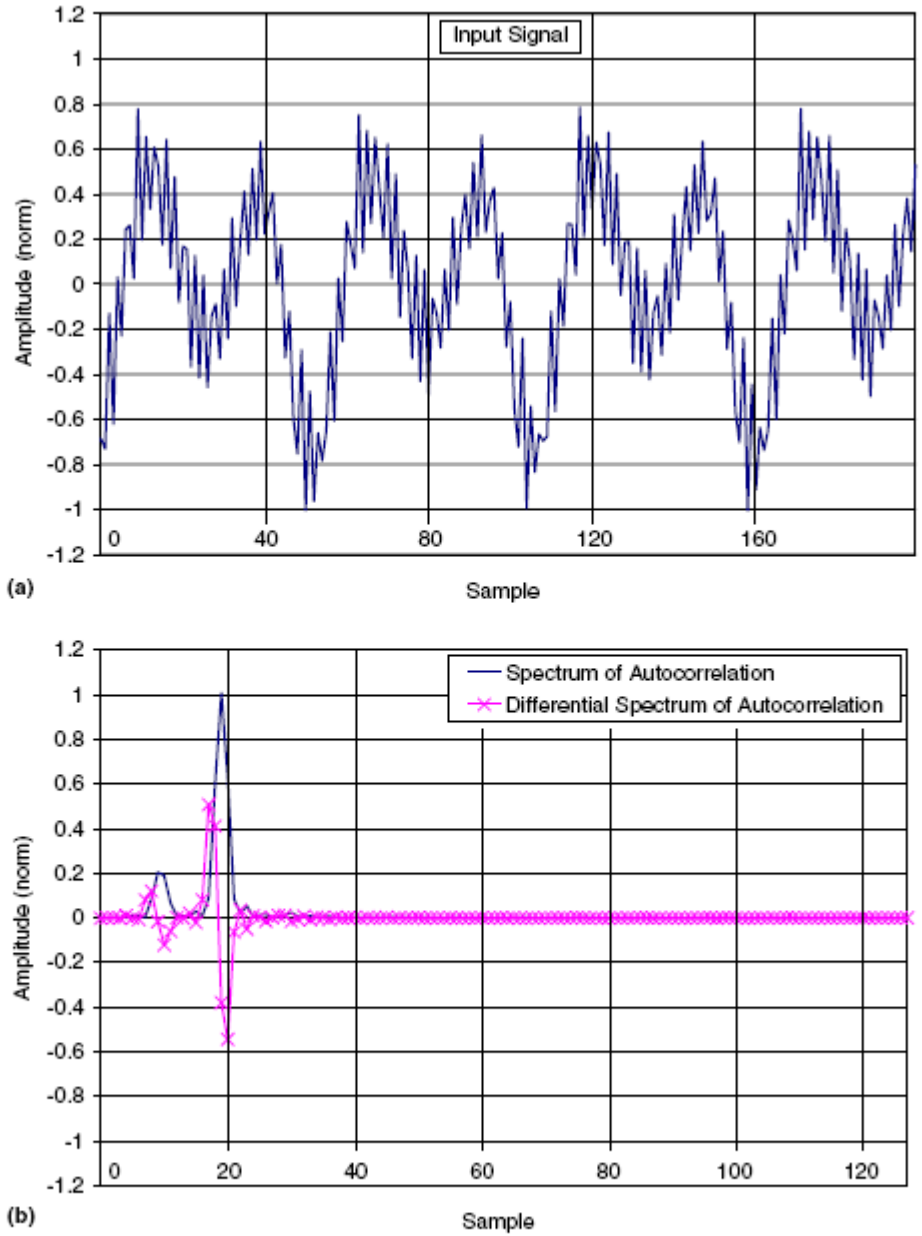


Fig. 1. (a) A sample speech signal, and (b) the autocorrelation spectrum magnitude and the differentiated autocorrelation spectrum magnitude of the same signal with a 512-point FFT. Only 128 points of the spectrum are shown for clarity. The sample signal is one frame of phone/iy/.

Experiments

- We tested our method on three different tasks, i.e., a small isolated word Farsi task, a medium-vocabulary continuous speech Farsi task and the noisy connected-digit Aurora 2 task.
- RAS: a filter length of L=2 will be used

$$H(z) = \frac{\sum_{t=-L}^{L} t z^{t}}{\sum_{t=-L}^{L} t^{2}} = \frac{-2z^{-2} - 1z^{-1} + z + 2z^{2}}{10}$$

- DPS :

$$Diff_Y(k) = Y(k) - Y(k+1)$$

Results on the isolated-word Farsi task

- **The speech corpus used in these experiments is a multi-speaker isolated-word Farsi(Persian) corpus.**

-The data were collected in normal office conditions with SNRs of 25 dB or higher and a sampling rate of 16 kHz.

- A total of 2665 utterance from 55 speakers were used for HMM model training.

- The test set contained 10 speakers (5 make and 5 female) that were not included in the training set.

-The noise then added to the speech in different SNR. The noise data were extracted from the NATO RSG-10 corpus(SPIB, 1995)

-We have consider babble, car, factory1 and white noise and added to the clean signal at 20, 15, 10, 5, 0 and -5 db SNRs.

Results on the isolated-word Farsi task

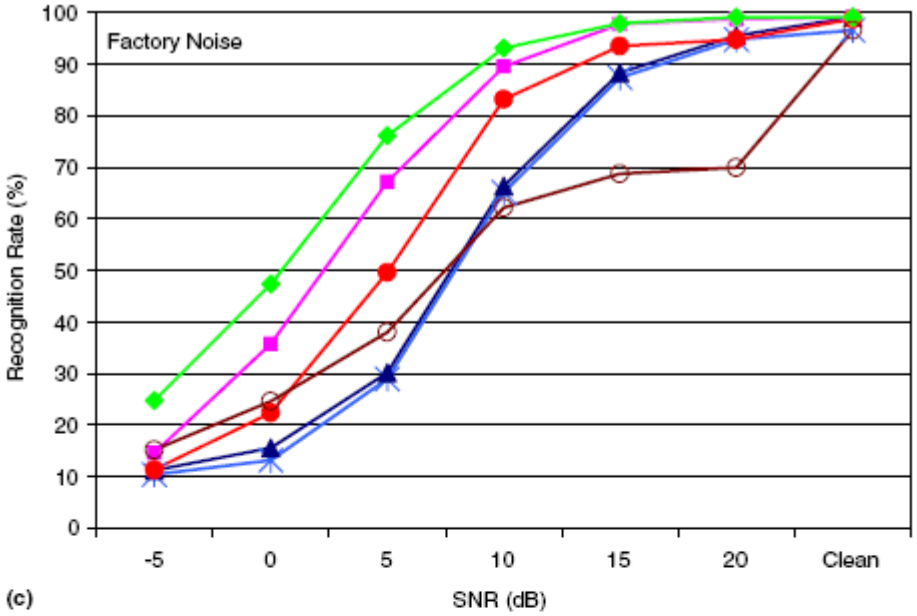
- **Our experiments were carried out using MFCC, MFCC applied to the signal enhance by spectral subtraction(SS), RAS-MFCC, DPS ,DAS.**

-The features in all cases were computed using 25ms frames with 10ms of frame shifts.

-Pre-emphasis coefficient was set to 0.97.-channel mel- scale filter-bank was used

-8-state left-right HMM and each state was represented by single-Gussian PDF.

-The feature vector were compose of 12 cepstral and log-energy parameters, together with their first and second-order derivatives(39 coefficients in total).



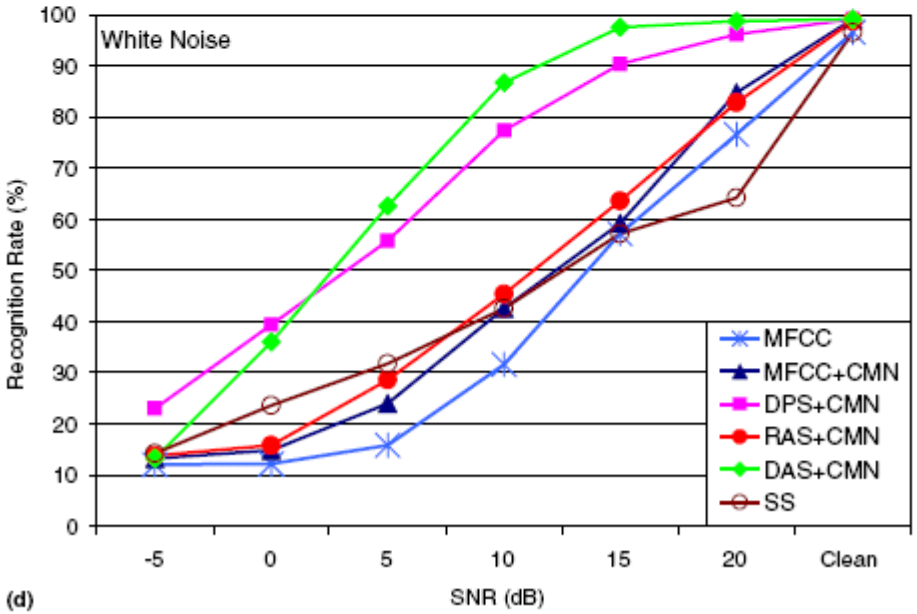


Fig. 3 (*continued*)

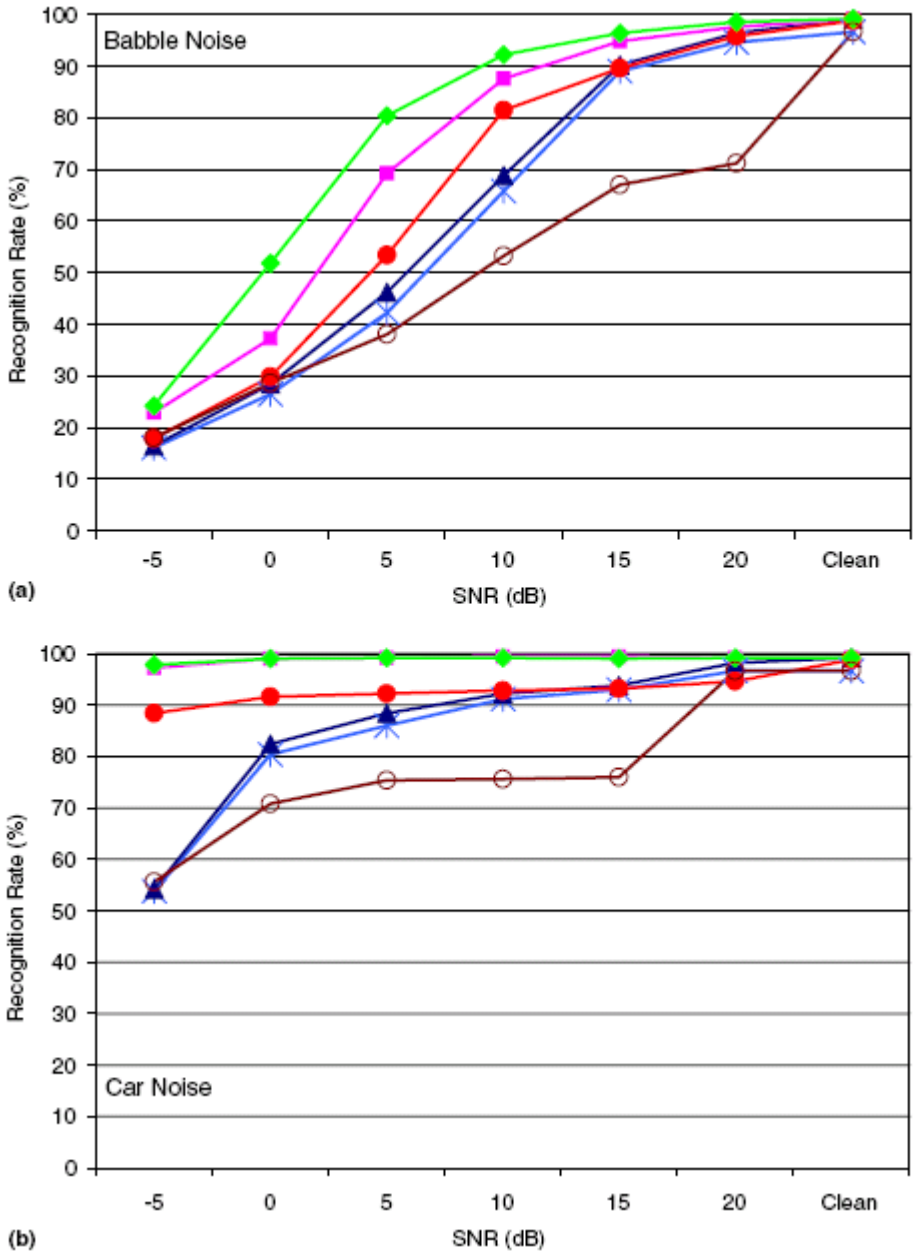

Babble Noise
Recognition Rate (%)
100
90
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
Clean
SNR (dB)
(a)
Car Noise
Recognition Rate (%)
SNR (dB)
(b)

Table 1
Comparison of baseline isolated-word 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 + CMN	98.80
DAS + CMN	99.20

Table 2
Comparison of average isolated-word recognition rates for various feature types with babble, car, factory and white noises

Feature type	Average recognition rate			
	Babble	Car	Factory1	White
MFCC (baseline)	63.6	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 + CMN	70.00 (10.06)	92.88(3.85)	68.72 (18.65)	47.24 (22.13)
DAS + CMN	83.88 (31.89)	99.12 (10.82)	82.80 (42.96)	76.36 (97.41)

Results on the contunuous Farsi task

- **The speech corpus used in these experiments is a speaker-independent medium-vocabulary continuous speech Farsi(Persian) corpus.**

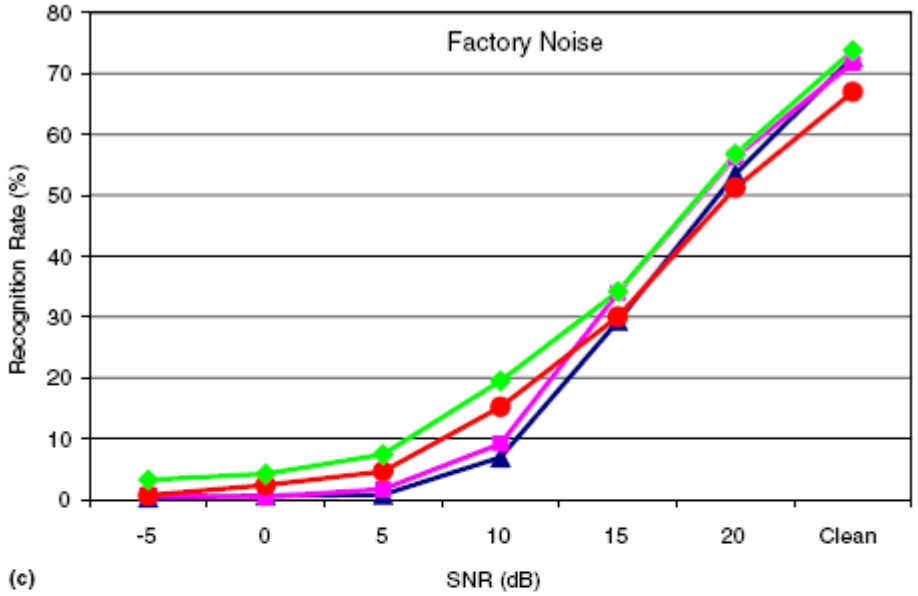
-The data were collected in normal office conditions with SNRs of 25 dB or higher and a sampling rate of 44.1 kHz.

-collect from 300 male and female adult speakers uttering 20 Persian sentences each, in two sessions.

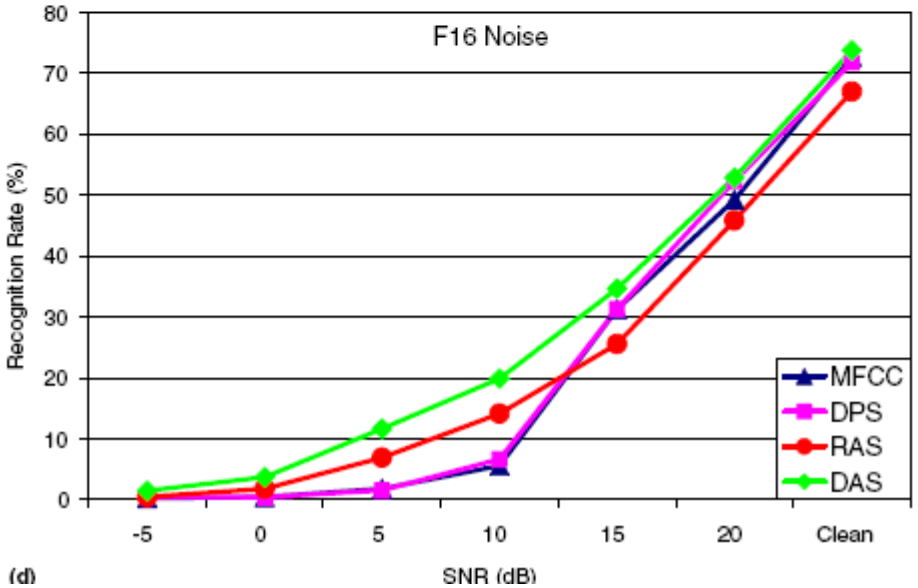
-The sentence uttered by each speaker were randomly selected from a set of around 400 sentences.

-The noise then added to the speech in different SNR. The noise data were extracted from the NATO RSG-10 corpus(SPIB, 1995)

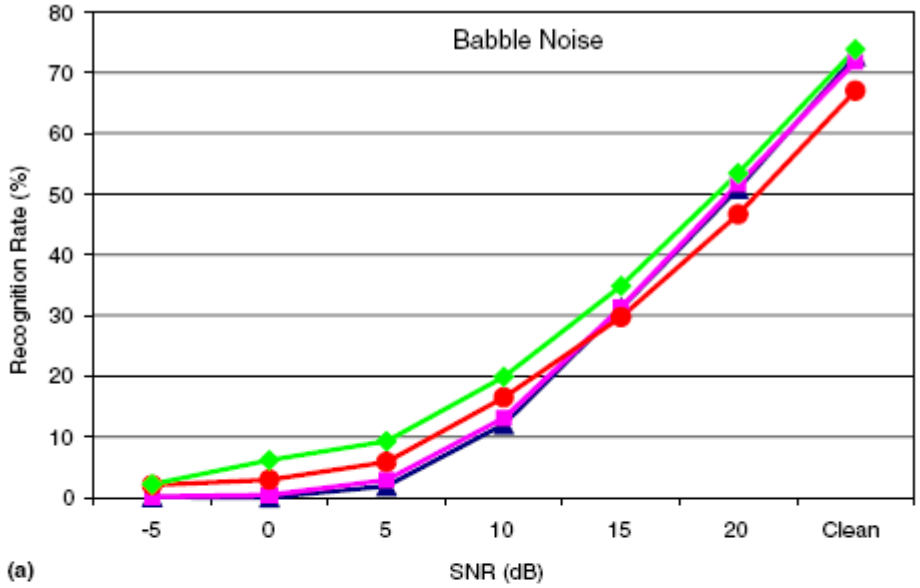
-We have consider babble, car, factory1 and white noise and added to the clean signal at 20, 15, 10, 5, 0 and -5 db SNRs.


Factory Noise
Recognition Rate (%)
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
Clean
SNR (dB)

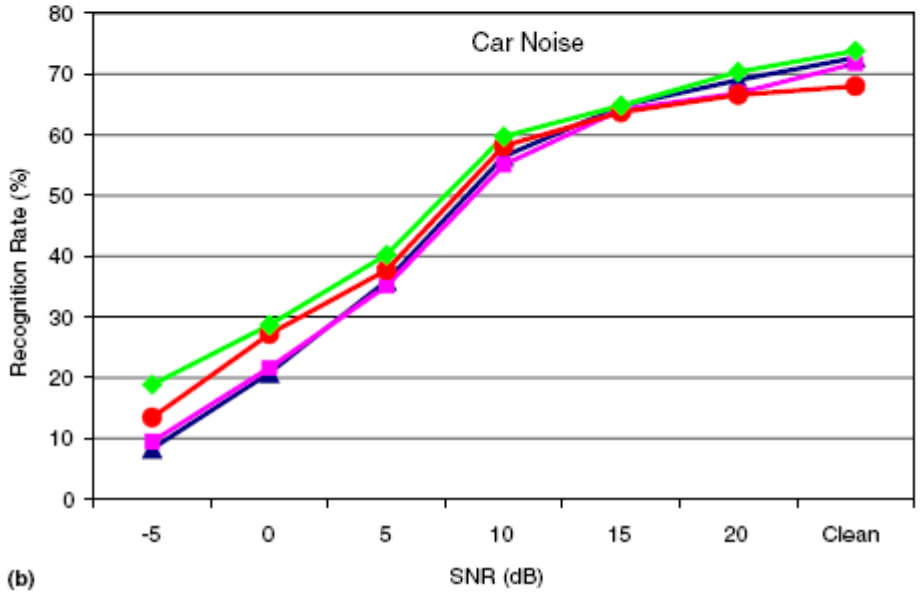
(c)


F16 Noise
Recognition Rate (%)
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
Clean
SNR (dB)
MFCC
DPS
RAS
DAS

(d)


Babble Noise
Recognition Rate (%)
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
Clean
SNR (dB)

(a)


Car Noise
Recognition Rate (%)
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
Clean
SNR (dB)

(b)

Table 3
Comparison of average continuous speech recognition accuracies for various feature types in babble, car, factory and f16 noises with different SNRs. Recognition was carried out using six mixture components per state

Recognition rate (%) – six mixture per state				
Noise type	Babble	Car	Factory	F16
MFCC (baseline)	19.23	49.39	18.23	17.64
MFCC + CMN	25.39 (32.03)	53.69 (8.71)	23.22 (27.37)	23.24 (31.75)
DPS	19.90 (3.48)	48.06 (−2.69)	20.32 (11.46)	18.42 (4.42)
DPS + CMN	26.43 (37.44)	53.62 (8.56)	23.8 (30.55)	23.54 (33.45)
RAS	20.32 (5.67)	50.67 (2.59)	20.7 (13.55)	21.84 (23.81)
RAS + CMN	26.44 (37.49)	54.21 (9.76)	25.05 (37.41)	24.03 (36.22)
DAS	24.72 (28.55)	52.78 (6.86)	26.22 (43.83)	26.42 (49.77)
DAS + CMN	31.31 (62.82)	57.09 (15.59)	28.67 (57.27)	28.84 (63.49)

Results on Aurora 2 task

- The aurora2 task includes two training modes, training on clean data only (clean-condition training)and training on clean and noisy data(muti-condition training).

-In clean-condition training, 8440 utterances from TIDigits containing those of 55 male and 55 female adults are used.

-For multi-condition mode, 8440 utterances from TIDigits training part are split equally into 20 subsets with 442 utterances in each subset

- Three test set are defined in Aurora 2, named A, B and C. 4004 utterance from TIDigits test data are divided into four subsets with 1001 utterance in each each subset at different SNRs.
- In the test set A, suburban train, babble, car and exhibition noise are added to the above mentioned four subsets.

- Test set B is created similar to test set A, but with four different noise, namely, restaurant, street, airport and train station.

- Test set C contains two of four subsets with speech and noise filtered using different filter characteristics in comparison to the data used in comparison to the data used in test seta A and B the noise used in this set are suburban train and street.

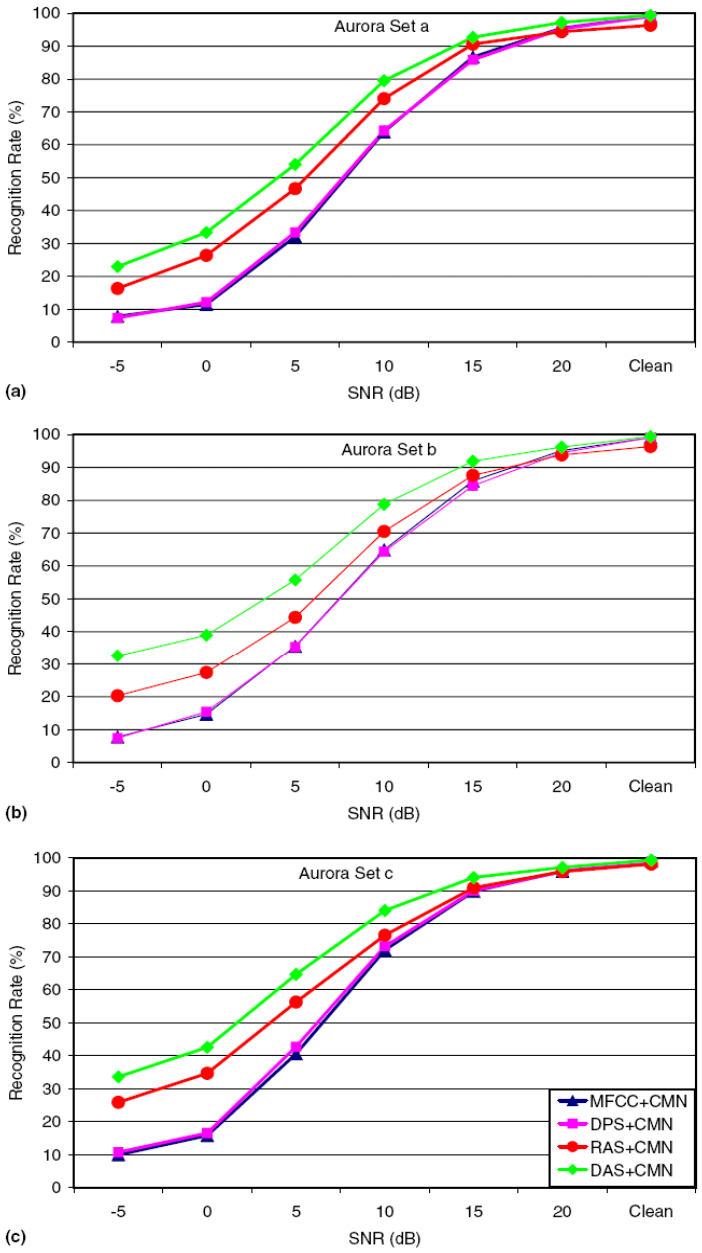


Fig. 6. Average recognition accuracies for clean-condition on AURORA2.0 (a) set A, (b) set B and (c) set C. The results correspond to MFCC, DPS, RAS and DAS front-ends with CMN.

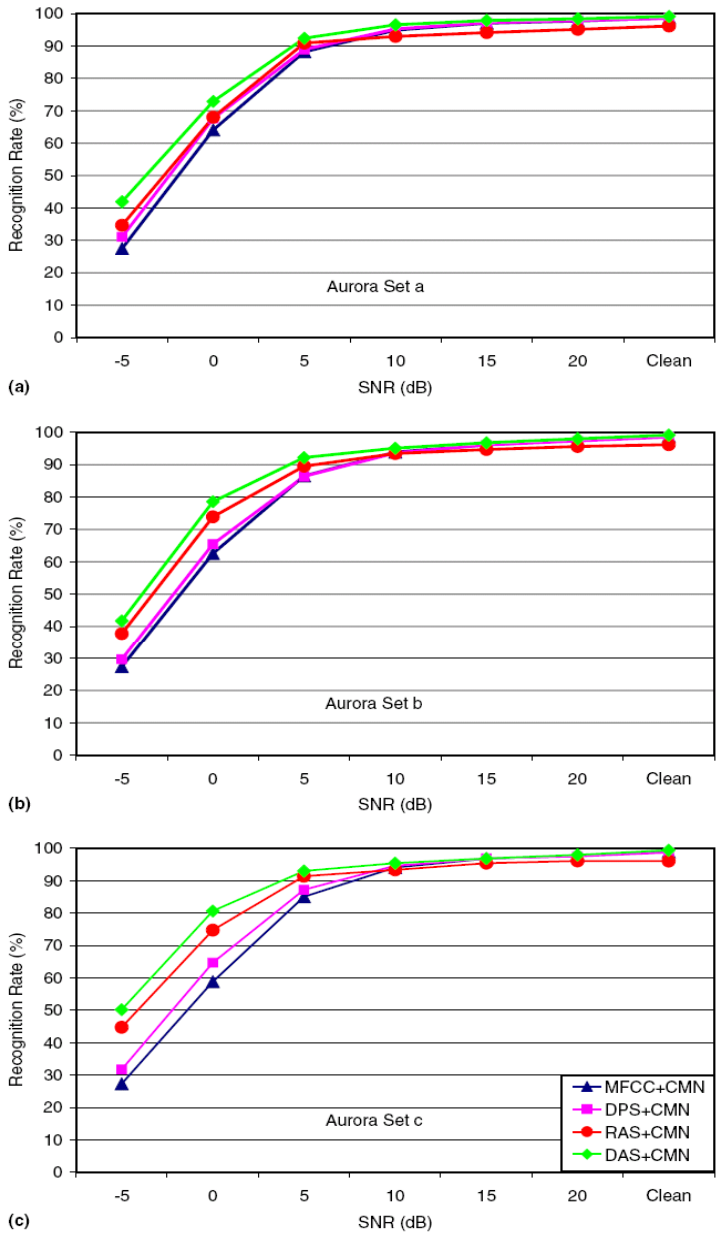


Fig. 7. Average recognition accuracies for multi-condition on AURORA2.0 (a) set A, (b) set B and (c) set C. The results correspond to MFCC, DPS, RAS and DAS front-ends with CMN.

Adjusting the parameters

- The parameters taken into consideration are the length of the FIR filter and the order of differentiation

$$H(z) = \frac{\sum_{t=-L}^{L} t z^{t}}{\sum_{t=-L}^{L} t^{2}} = \frac{-2z^{-2} - 1z^{-1} + z + 2z^{2}}{10}$$

$$Diff_Y(k) = Y(k) - Y(k+1)$$

$$Diff_Y(k) = Y(k) - Y(k+2)$$

$$Diff_Y(k) = Y(k-2) + Y(k-1) - Y(k+1) - Y(k+2)$$

Table 4
The average continuous speech recognition rate for various noise and SNRs with different filter lengths. Biased estimator was used for one-sided autocorrelation sequence and the models featured six mixture components per state

DAS features				
Filter length	Noise type			
	Babble	Car	Factory	F16
$L = 1$, 3 frames	22.07	49.06	22.59	21.72
$L = 2$, 5 frames	22.97	50.00	23.24	22.95
$L = 3$, 7 frames	24.69	51.34	24.41	24.42
$L = 4$, 9 frames	23.82	50.89	23.76	23.15
$L = 5$, 11 frames	23.07	49.91	23.35	22.89

Table 5
The average continuous speech recognition rate for various noise and SNRs with different filter lengths

DAS features				
Filter length	Noise type			
	Babble	Car	Factory	F16
$L = 1$, 3 frames	24.08	51.8	23.77	23.25
$L = 2$, 5 frames	24.72	52.78	24.43	24.53
$L = 3$, 7 frames	26.00	53.24	26.49	25.38
$L = 4$, 9 frames	25.31	53.06	25.61	23.78
$L = 5$, 11 frames	25.09	52.59	25.41	23.76

Unbiased estimator used for one-sided autocorrelation sequence and the models featured six mixture components per state.

Table 6
CSR averaged accuracies over various SNRs with different noise types, filter length $L = 3$ and unbiased estimator for one-sided autocorrelation sequence obtained using (15)

Noise type	Recognition rate
Babble	24.83
Car	51.56
Factory	25.17
F16	23.71

Table 7
CSR averaged accuracies over various SNRs with different noise types, filter length $L = 3$ and unbiased estimator for one-sided autocorrelation sequence obtained using (16)

Noise type	Recognition rate
Babble	25.19
Car	52.35
Factory	25.82
F16	24.25