SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effect on MFCC feature

Author: Babak Nasersharif*, Ahmad Akbari

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Compensating of noise effects on MFCC features
- Mel sub-band spectral subtraction
- SNR-dependent compression of Mel sub-band energies
- Experiments and results

Introduction

- Current automatic speech recognition (ASR) systems are not robust in adverse acoustic conditions.

  -back ground noise .

  -channel distortion.

  -unwanted sounds.

- For these reason, there is demand for techniques to compensate for the effects of such interfering signals.

Introduction(cont.)

- Several technique have been proposed to reduce sensitivity of feature to external noise.

- A group of methods work at spectral level.
  - Spectral subtraction.
  -Wiener filtering.

- Another robust speech recognition techniques work at feature level.
  - cepstral mean normalization.(CMN)
  - SNR-dependent cepstral mean normalization.(SDCMN)

Introduction(cont.)

- In this paper, we propose a transformation for applying to Mel sub-bands energies in order to remove noise from MFCC feature.

-First step, we apply sub-band spectral subtraction.

-second step, we define an SNR-dependent root function in place of log function and we use it for compressing Mel sub-band energies.

Compensating of noise effects on MFCC features

- To overcome MFCC have poor performance in noisy condition, we propose a framework to compensate additive noise.

- First discuss the general process of MFCC feature extraction

  -assume that $x(n)$ represents a speech signal .

  -pre-emphasize.

  -multiplied by a Hamming window with length N.

  -applying an N-point fast Fourier transform(FFT).

  -the resulting amplitude spectrum is shown by $|X(k)|$, where k is frequency index.

Compensating of noise effects on MFCC features(cont.)

- Then the filter bank energy E_i^x passing through ith Mel band-pass filter $\psi_i(k)$, is calculate as follows:

$$E_i^x = \sum_{k=1}^{N} |X(k)|^2 \cdot \psi_i(k)$$

- After, a discrete cosine transform(DCT) is applied to log of filter bank energies.

$$c_t^x = \sum_{i=1}^{M} \log(E_i^x) \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

Compensating of noise effects on MFCC features(cont.)

- Assuming that x(n) is noisy speech, we define our noise compensation framework based on filter bank energies.

$$\hat{E}_i^x = F(E_i^x, w_i, b_i) = (E_i^x - b_i)^{w_i}$$

where $\hat{E}_i^X$ is compensated Mel filter bank output , w_i and b_i are compensation parameters. The parameter w_i is the compression factor and the bias b_i depends on noise spectral characteristics.

- Including two steps:

-Subtraction: reduce the filter bank energy increase due to present of additive noise.

-Energy compression: emphasize those filter bank energies less affected by noise and distortion.

Compensating of noise effects on MFCC features(cont.)

- After these two steps, we can calculate the compensated MFCC using the following equation.

$$\hat{c}_l^x = \sum_{i=1}^{M} \hat{E}_i^x \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$
$$= \sum_{i=1}^{M} (E_i^x - b_i)^{w_i} \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

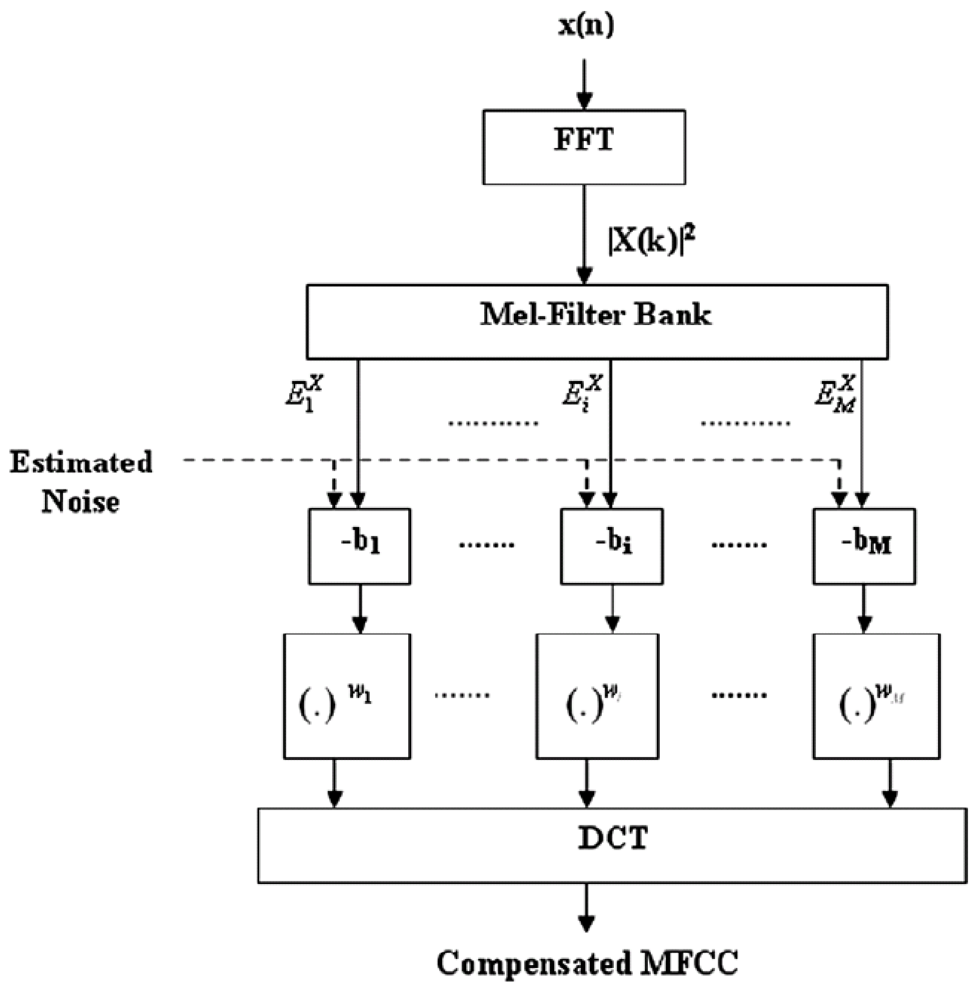


Fig. 1. Block diagram of proposed method for compensation of noise effects on MFCC features.

Mel sub-band spectral subtraction

- Conventional power spectral subtraction is defined as follows:

$$|\hat{S}(k)|^2 = \begin{cases} |X(k)|^2 - \alpha |N(k)|^2 & \text{if } |X(k)|^2 > \frac{\alpha}{1-\beta}|N(k)|^2 \\ \beta |S(k)|^2 & \text{otherwise} \end{cases}$$

Where $|\hat{S}(k)|^2$, $|X(k)|^2$ and $|N(k)|^2$ are the power spectral of enhanced speech, noisy speech, and estimate noise. α is an over-estimation factor and β is a spectral flooring parameter.

- In this paper, we use Mel sub-band spectral subtraction

$$E_i^{ss} = E_i^x - b_i = \begin{cases} E_i^x - \alpha_i E_i^N & E_i^x > \frac{\alpha_i}{1-\beta_i} E_i^N \\ \beta_i E_i^x & \text{otherwise} \end{cases}$$

where E_i^{ss} is enhanced filter bank energy after Mel sub-band spectral subtraction. E_i^N is the output of ith triangular Mel-scaled filter, assuming that estimated noise $|N(k)|^2$ is passed through Mel filter bank.

Mel sub-band spectral subtraction(cont.)

- E_i^N can be computed as follows:

$$E_i^N = \sum_{k=1}^{N} | N(k) |^2 \cdot \psi_i(k)$$

- And we can compute parameter

$$b_i = \begin{cases} \alpha_i E_i^N & E_i^x > \dfrac{\alpha_i}{1-\beta_i} E_i^N \\ (1-\beta_i) E_i^x & otherwise \end{cases}$$

Mel sub-band spectral subtraction(cont.)

- We estimate the noise power spectrum using 300ms of noisy speech signal where only the noise is present.

- We use following smoothing equation for the noise power spectrum estimation.

$$|N(k)|^2 = P_t(k) = \lambda P_{t-1}(k) + (1-\lambda)|B_t(k)|^2$$

where $P_{t-1}(k)$ and $|B_t(k)|^2$ are estimated noise power spectral in previous t-1 frames and current frame.
λ is a forgetting factor and k is the frequency index.

SNR-dependent compression of Mel sub-band energies

- For MFCC computation, a logarithm function applied to Mel filter bank energies in order to compress their dynamic range.

- This reduction has two drawbacks in presence of addictive noise.

  - It can not highlight sub-bands energies that are less affected by noise.
  - some distortions that ate negligible in power spectrum domain become important after the logarithmic compression of Mel filter bank energies.

SNR-dependent compression of Mel sub-band energies(cont.)

- DCT that is utilized in MFCC computation is a linear transform that gives equal weight to all compressed sub-band energies.

- Equal weight of DCT and drawbacks of logarithmic compression make MFCC feature highly sensitive to additive noise.

SNR-dependent compression of Mel sub-band energies(cont.)

- We propose a compression function that is computed based on SNR in Mel sub-bands. This function replaces w_i in before equation.

$$w_i = \gamma \cdot \left[1 - \exp(-\frac{SNR_i}{\xi_i}) \right] = \gamma \cdot G(SNR_i, \xi_i)$$

where γ is a constant root and ξ_i is a parameter that controls the steepness of the compression function. G is an SNR-dependent function with values between 0 and 1.

- SNR_i is signal to noise ration ith Mel frequency sub-band that can be estimated as in:

$$SNR_i = (1 + \frac{E_i^{ss}}{E_i^N})^{0.5}$$

where square root has been used for reducing the dynamic rage of energy ration.

SNR-dependent compression of Mel sub-band energies(cont.)

- We need more compression at sub-bands with low SNR_i value while we need less compression or equalization at sub bands with high SNR_i value.

$$\xi = 1 - \frac{1}{1+\exp\left(-\frac{SNR_i - \mu SNR}{\sigma SNR}\right)} = 1 - f(SNR_i)$$

where μ_{SNR} and ρ_{SNR} are mean and standard deviation of SNR_i computed from all Mel sub-bands of a speech frame.

- Function f was chosen as a sigmoid function, because a sigmoid function f satisfies asymptotic behavior of being zero at low SNR_i and one at high SNR_i.

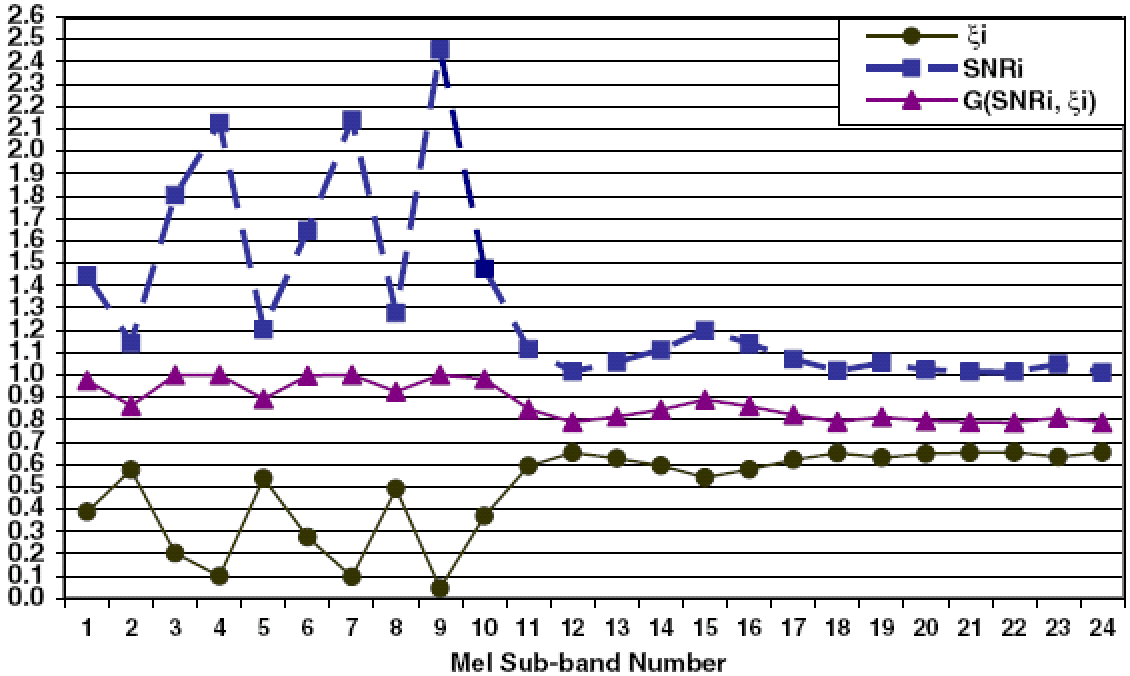


Fig. 2. SNR_i, ξ_i and $G(SNR_i, \xi_i)$ values in different Mel sub-bands for a noisy speech frame in presence of white noise with SNR = 0 dB.

Experiments and results

- TIMIT database for isolated word recognition.

  -two sentences spoken by speakers from two dialect regions were selected and were segmented into words.

  - had 21 words spoken by 151 speakers including 49 females and 102 males.

  -training set contains 2349 utterances spoken by 114 speakers.

  -the testing set including 777 utterances spoken by 37 speakers.

  -recognizer is CDHMM with 6 states and 8 Gaussian mixture per state.

- Three type of additive noises were used: white, pink, factory noise selected from NOISEX92 database.

Experiments and results(cont.)

- For evaluating our proposed compensation method, we have tested Mel sub-band spectral subtraction in company with conventional log function.

$$sc_t^x = \sum_{i=1}^{M} \log(E_i^{ss}) \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

- Denote this features by LMSBS.

Experiments and results(cont.)

- CMSBS which stands for Compression and Mel Sub-Band Spectral subtraction show our proposed features. We have chosen α_i =1, β_i =0.1 for all Mel sub-bands .

$$\hat{c}_t^x = \sum_{i=1}^{M} (E_i^{ss})^{w_i} \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

Experiments and results(cont.)

- Moreover, we compare CMSBS with constant root where we choose the constant root equal to 0.5

$$rc_t^x = \sum_{i=1}^{M} (E_i^X)^{0.5} \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

- Denote this features by RMFCC .

Experiments and results(cont.)

- Furthermore, we have performed Mel spectral subtraction together with constant root 0.5 that can be shown by:

$$src_t^x = \sum_{i=1}^{M} (E_i^{ss})^{0.5} \cos[l \cdot \frac{(2i-1)\pi}{2M}]$$

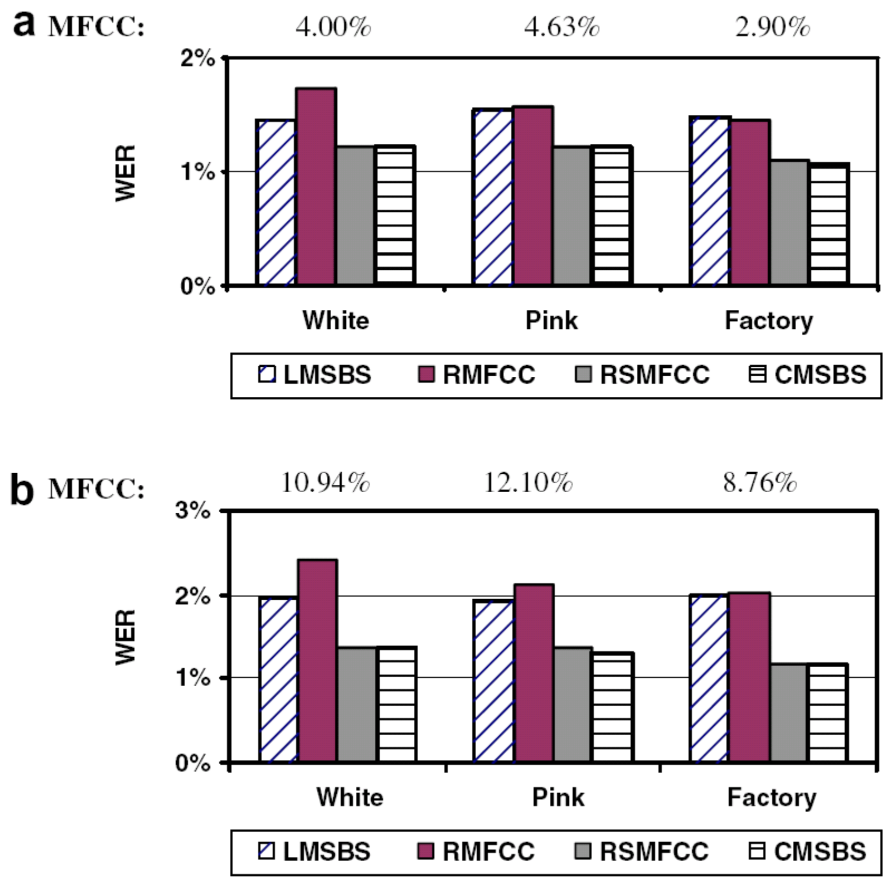
- Denote this features by RSMFCC.

Experiments and results(cont.)

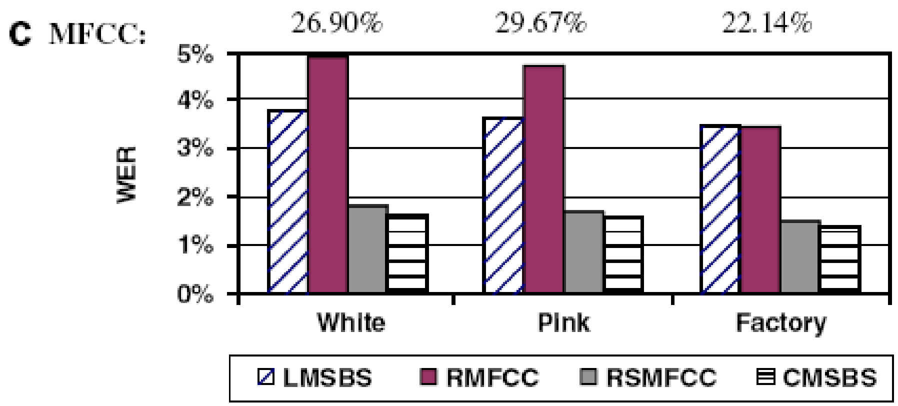
Table 1
Word error rates for conventional MFCC features in presence of white, pink and factory noises for different SNR values

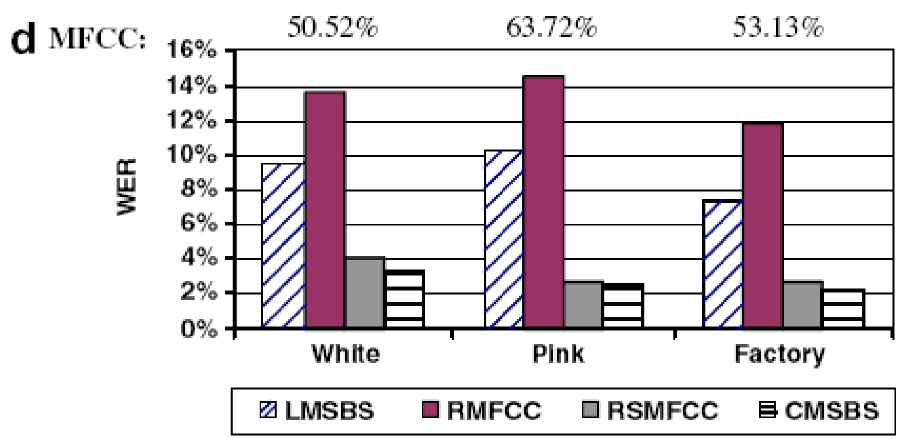
Noise	SNR				
	20 dB	15 dB	10 dB	5 dB	0 dB
White (%)	4.00	10.94	26.90	50.52	81.39
Pink (%)	4.63	12.10	29.67	63.72	90
Factory (%)	2.90	8.76	22.14	53.13	85.84

Experiments and results(cont.)



Experiments and results(cont.)





Experiments and results(cont.)

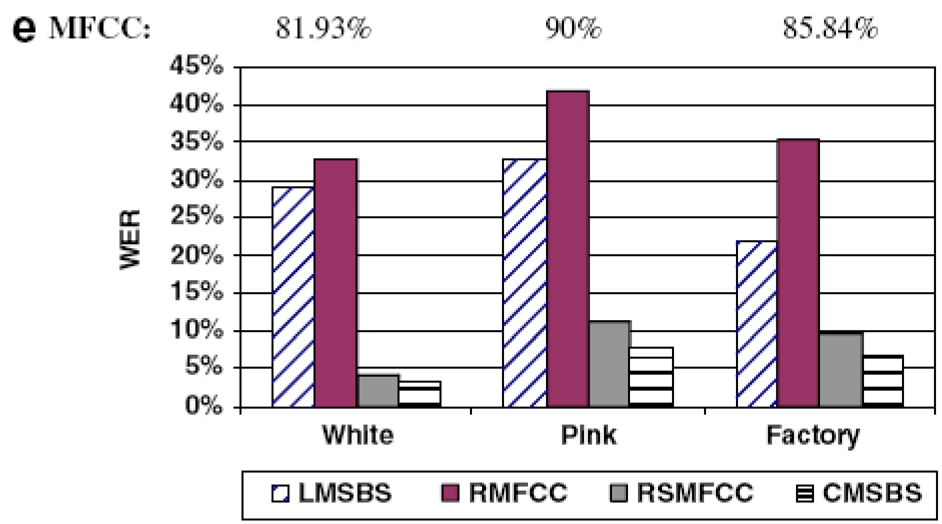
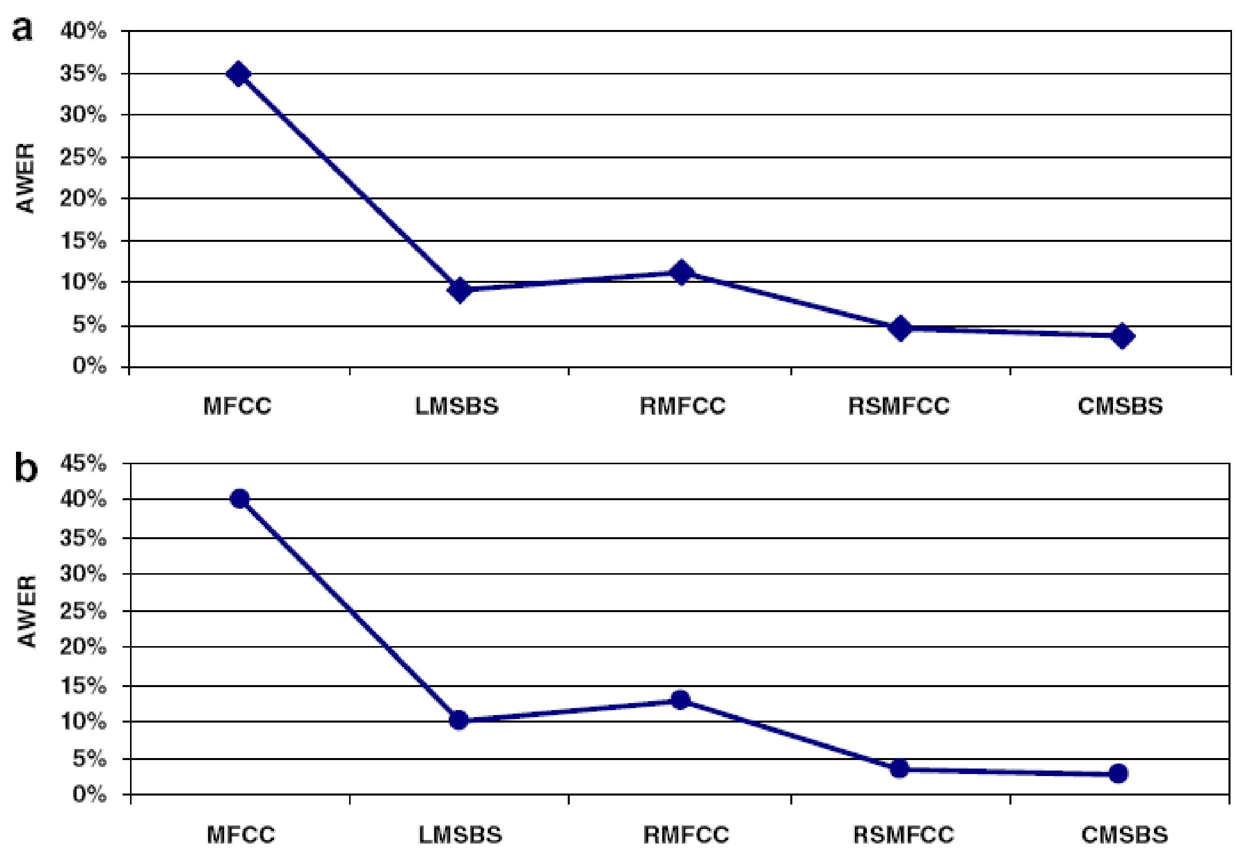


Fig. 3. Word error rates in presence of white, pink and factory noises for different SNR values: (a) SNR = 20 dB, (b) SNR = 15 dB, (c) SNR = 10 dB, (d) SNR = 5 dB, (e) SNR = 0 dB.

Experiments and results(cont.)

B. Nasersharif, A. Akbari / Pattern Recognition Letters 28 (2007) 1320–1326



Experiments and results(cont.)

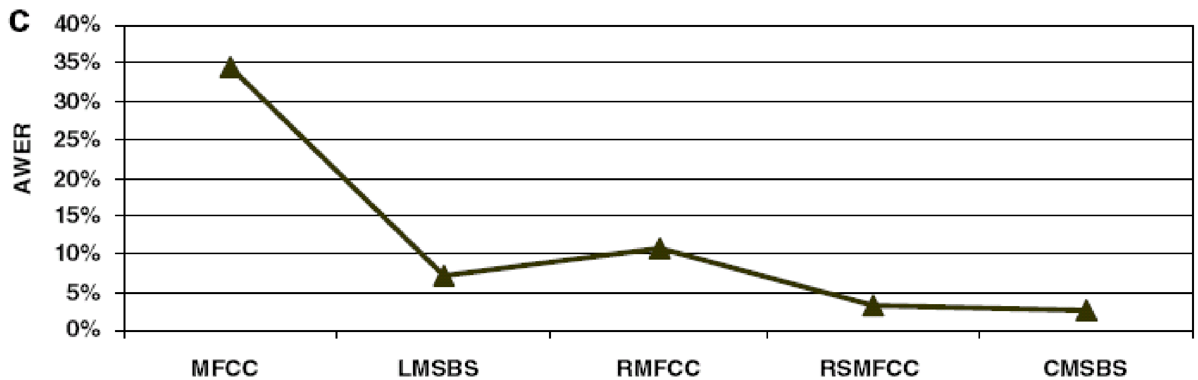


Fig. 4. Average word error rate on five SNR values (20, 15, 10, 5, 0 dB) for three noise types: (a) white noise, (b) pink noise, (c) factory noise.