Improved MFCC-Based Feature for Robust Speaker Identification

Author : Zunjing WU, Zhigang CAO
Professor: 陳嘉平
Reporter: 吳國豪

Outline

- Introduction
- Compression Function
- Combining Proposed Function with SS and MF
- Experiment And Results

Introduction

- The techniques for robust speaker recognition can be classified into three categories based on the working spaces.

(1) Pre-processes the noisy speech signal to obtain a better estimation of clean speech.

(2) Focuses on robust feature representation of the speech.

(3) Acoustic model parameters are adapted to match the noisy speech.

Introduction

- The log function in the MFCC generation is very sensitive to noise.

(1) The spectral subtraction(SS) method effectively suppresses noise to improve the signal-to-noise (SNR) of the input speech.

(2) A median-filter was used to smooth the filter bank energies and to restrain the high-frequency components.

(3) Using a power function to replace the log function.

Compression Function

- The standard MFCC analysis consists of five steps:

1) Pre-process the input speech and detect the endpoints.
2) Perform an fast Fourier transform on the input speech signal.
3) Calculate the Mel-frequency bank energies by integrating the spectral energy coefficients within triangular frequency bins arranged uniformly on the Mel-frequency scale.
4) Perform the discrete cosine transform on the logarithm of the filter-bank energies.
5) Append first order differentials.

Compression Function

- In step 3, the log transformation nonlinearly compresses the filter-bank energies in accord with a human auditory response. However, the log transformation is sensitive to noise because its slope is very steep when the energy is low, so serious mismatches are introduced in low-energy log filter-banks.

- In Fig. 1, noise influence on the log filter-banks at low energies is much larger than high-energy banks due to the log function effect.

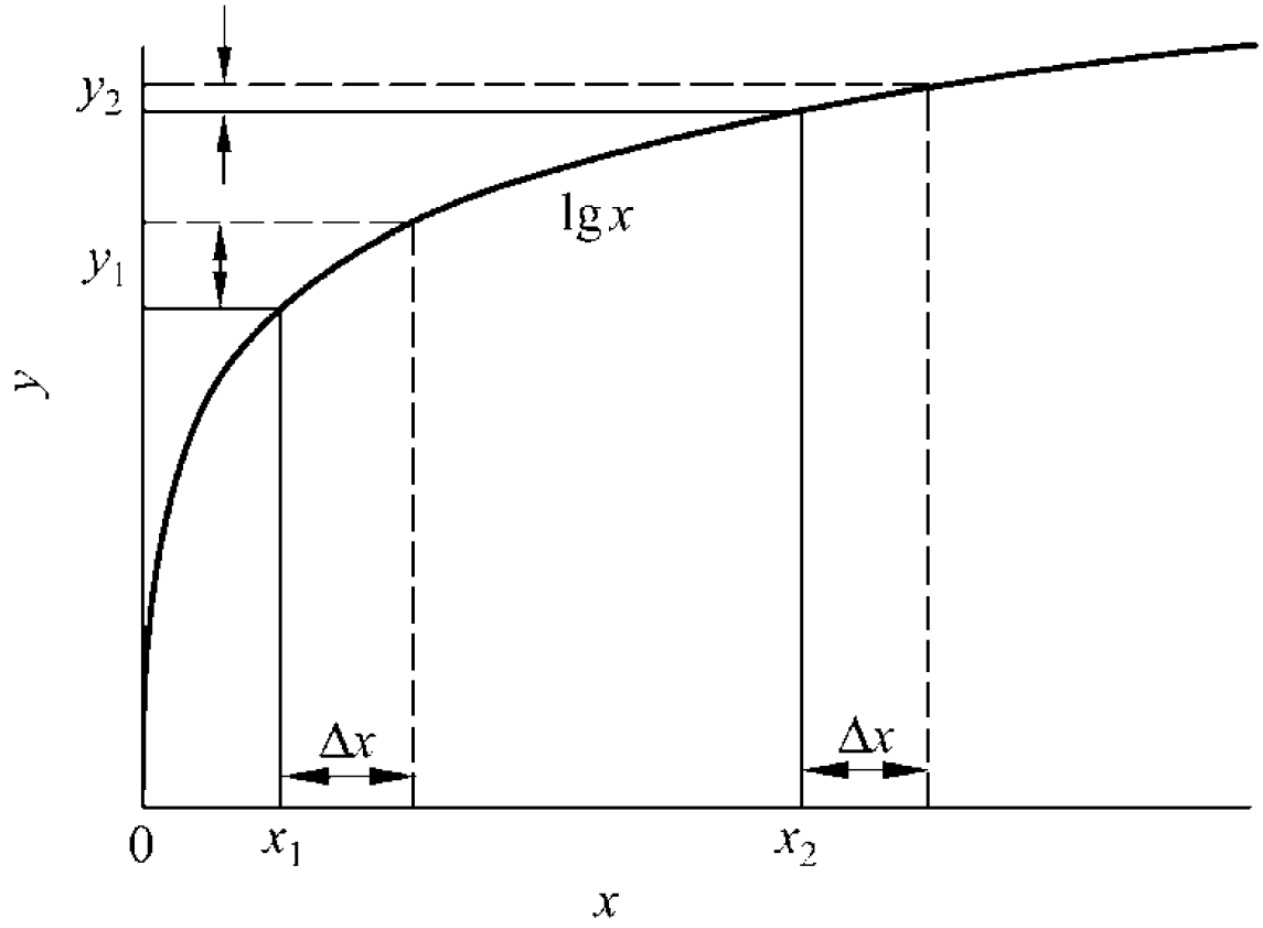

Fig. 1 Mismatches induced by log transformation

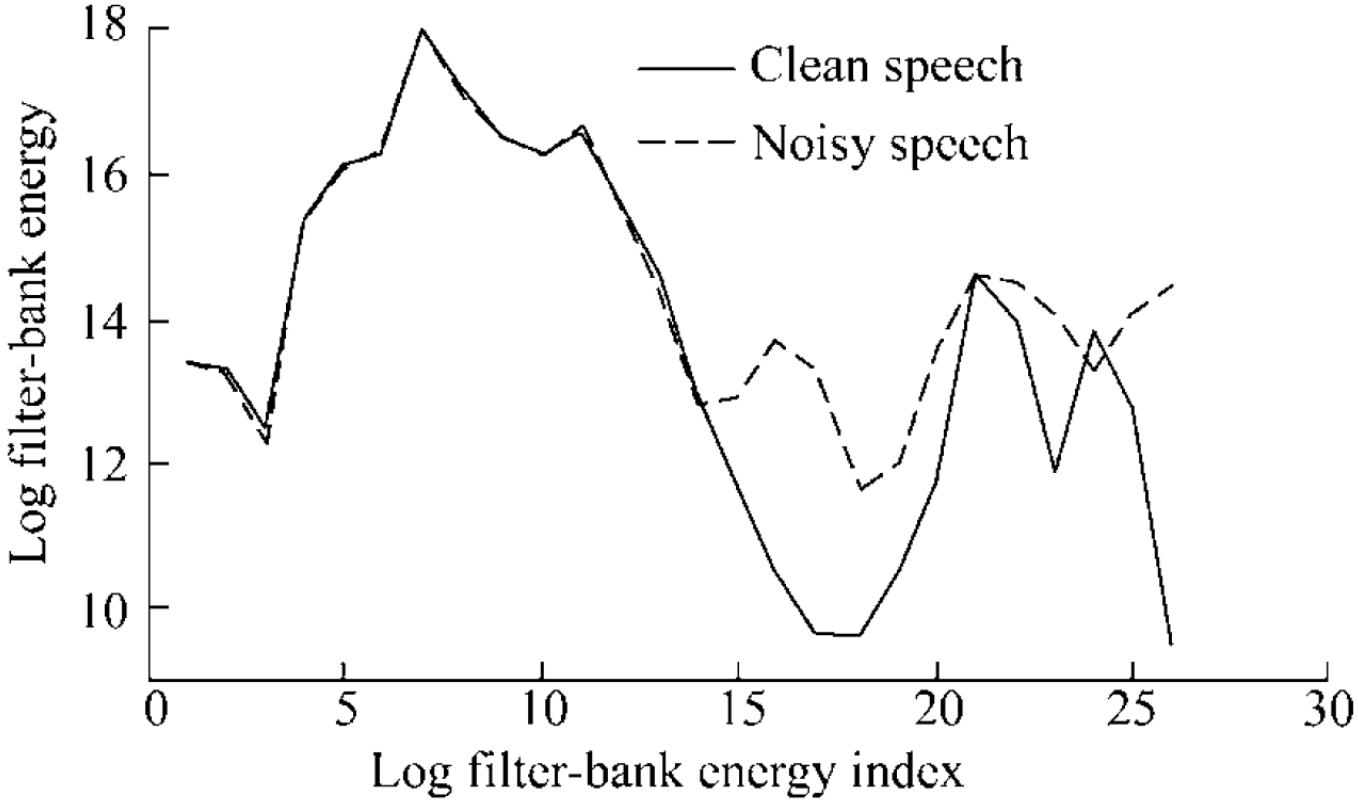

Fig. 2 Log filter-bank energies of clean and noisy speech

Compression Function

- The lower segment of the log function was replaced by a power function as:

$$f_{\mathrm{PL}}(x) = \begin{cases} \lambda x^{\frac{1}{\lambda}} / C^{\frac{1}{\lambda}} & x \le C; \\ \lg x + \lambda - \lg C, & x > C. \end{cases}$$

C is the noise masking level and λ is the compression coefficient.

- The filterbank energies using the combined transformation with C= 5×10^6 and λ=2 are shown in Fig. 3. The mismatch is much less in Fig. 3 for low-energy banks.

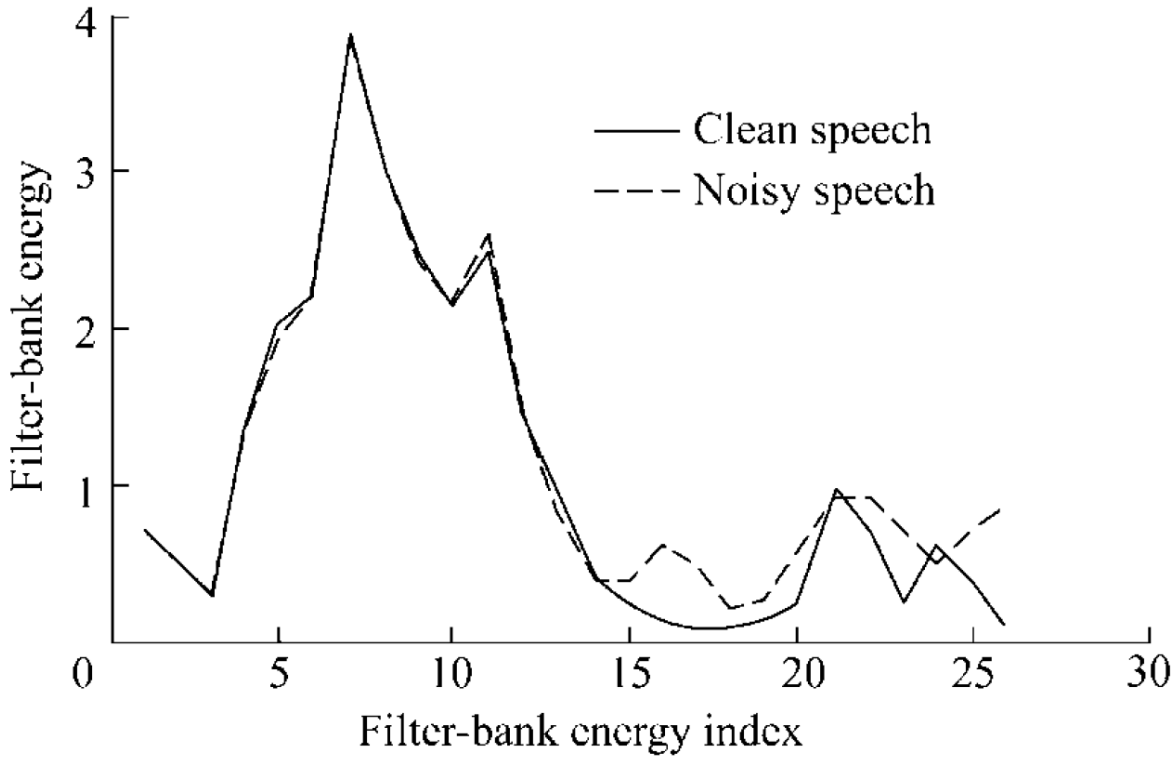

Fig. 3 Filter-bank energies using the combined transformation

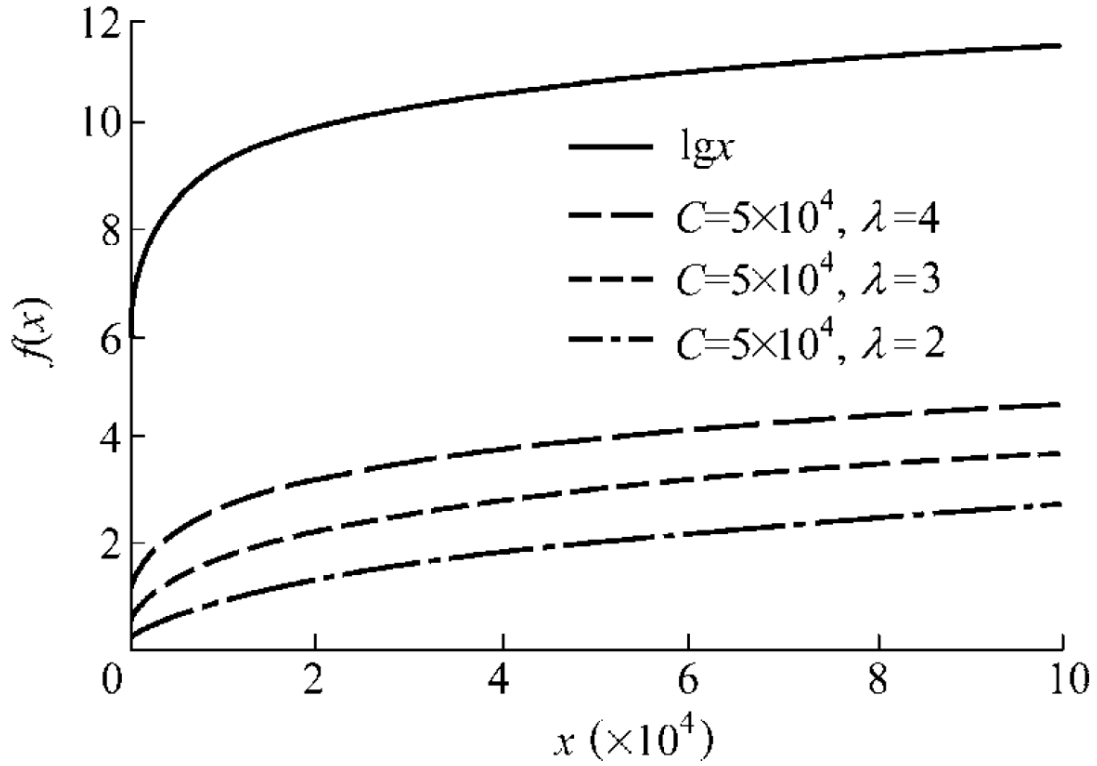

Fig. 4 Transformation functions corresponding to different parameters *C* and *λ*

Combining Proposed Function with SS and MF

- In step 1, adding the SS method.
- In step 3, a median-filter was used to smooth the filter bank energies and to restrain the high-frequency components.
- In step 4 used the transformation function $f_{pl}(x)$ to replace the log function.

Experiment

- The system was evaluated with a 26-dimensional MFCC_0_D and a 32-mixture Gaussian mixture model.
- A set of 30 speakers was selected from the dr6' region of the TIMIT database, including 16 females and 14 males.
- Noise was added to test sentences artificially at different SNR levels of 0 dB, 5 dB, 10 dB, 20 dB, and clean speech.

Experiment 1

- The first test evaluating the performances of the baseline system and enhanced systems using the SS and SS_MF (combination of SS and MF) methods.

- The results in Fig. 5 show that the baseline system performance degraded rapidly with the SS method providing improved system performance in noisy environments and the SS_MF method providing better performance in the low SNR ranges.

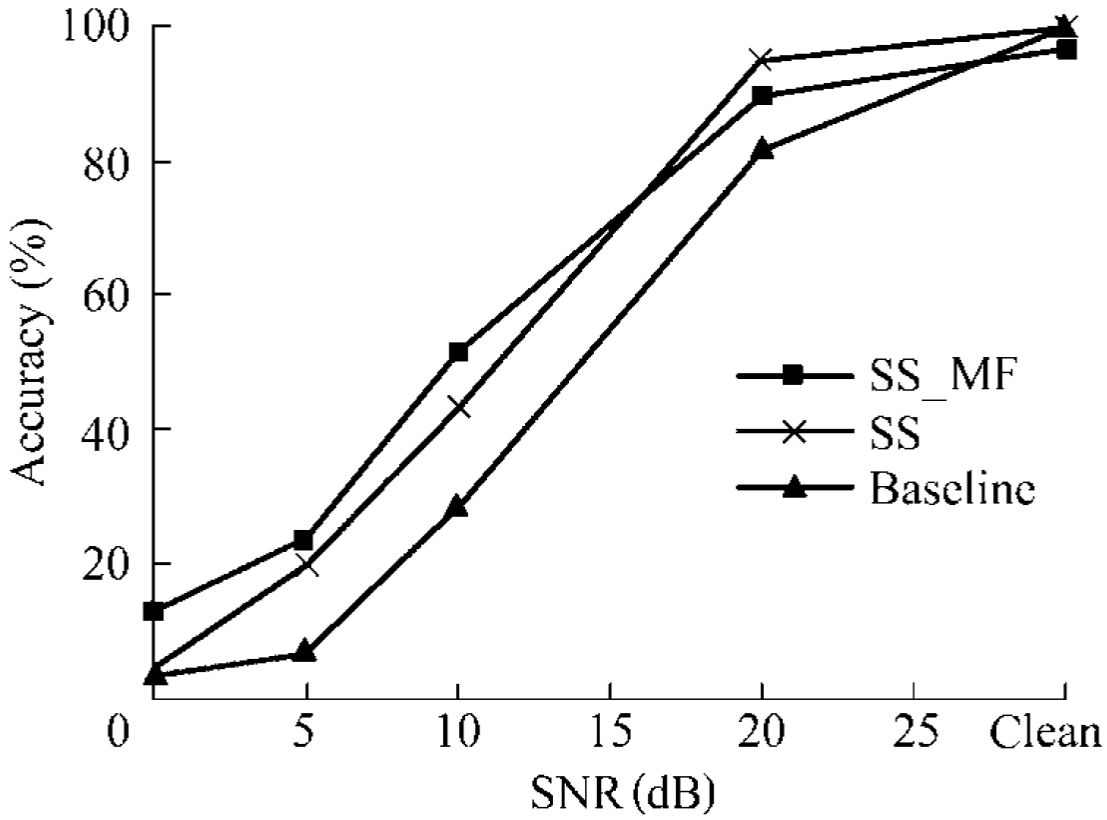

Fig. 5 Performance of baseline, SS, and SS_MF methods

Experiment 2

- The second set of tests evaluated the combined SS_PL_MF method with various values of C and λ.
- The results in Figs. 6 and 7 show that the present improved system very effectively improves the system performance in very noisy environments compared with the SS methods.
- Increasing the value of C increases the identification accuracy for low SNR while degrades the identification accuracy for high SNR because of the competing influences of the lost of speech information and the noise reduction.

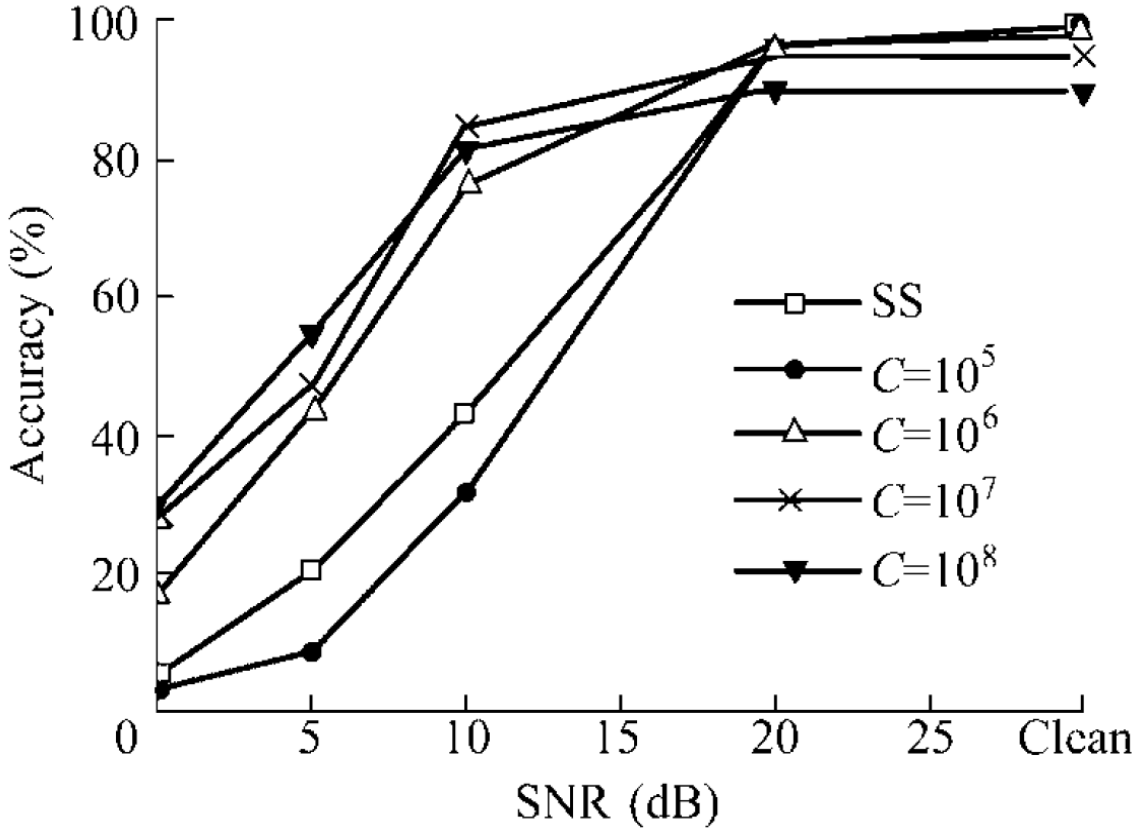

Fig. 6 Performance of SS_PL_MF system for λ=2

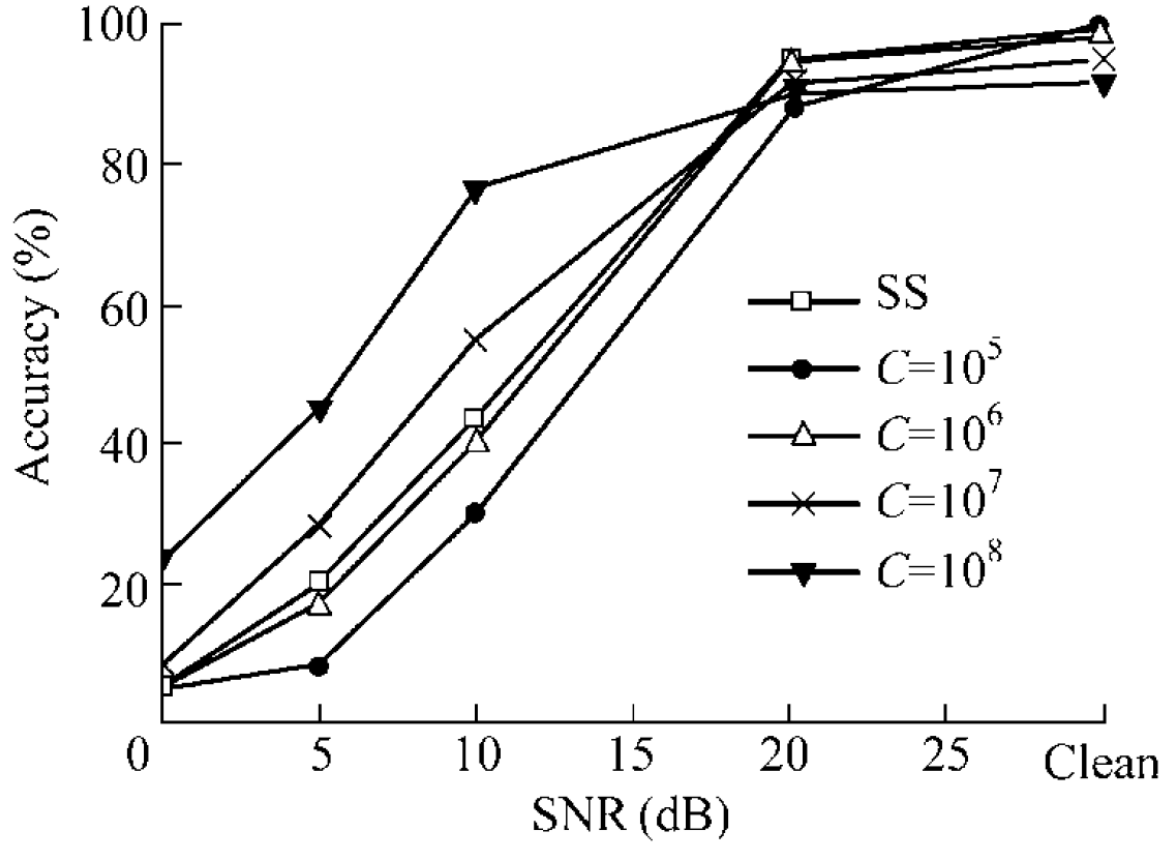

Fig. 7 Performance of SS_PL_MF system for λ =3

Experiment 2

- In general, the combined system with λ =2 performs better than with λ =3.
- For λ =2,C= 10^8 achieved the best performance for 0 dB and 5 dB, while C= 10^7 was better for 10 dB and C= 10^6 was better for 20 dB.
- The system performance with C=10^7 and λ=2 was close to the performance of the optimal system using different C for different noise levels.