Increased MFCC Filter Bandwidth For Noise-Robust Phoneme Recognition

Author :Mark D.Showronski ,

John G.Harris

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Filter Widening Schemes
- Experiment And Conclusion

Introduction

- Many speech recognition systems use mel-frequency coefficient(mfcc) feature extraction as a front-end. In the algorithm, a speech spectrum passes through a filter bank of mel-spaced triangular filters, and the filter output energies are log-compressed and transformed to the cepstral domain by the DCT.

- With complex cochlear models of human auditory systems, the filters in the model's filter bandwidth are much wider and overlap with neighboring filters more so than mfcc filters.

Filter Widening Schemes(1)

- The first scheme for widening the mfcc filters increase this overlap while maintaining the bandwidth of entire filter bank. Thus the triangle base length L for each filter is:

$$L=\frac{\hat{f}_{max}-\hat{f}_{min}}{N(1-m)+m}$$

$\hat{f}_{max}$: the maximum frequency of the filter bank

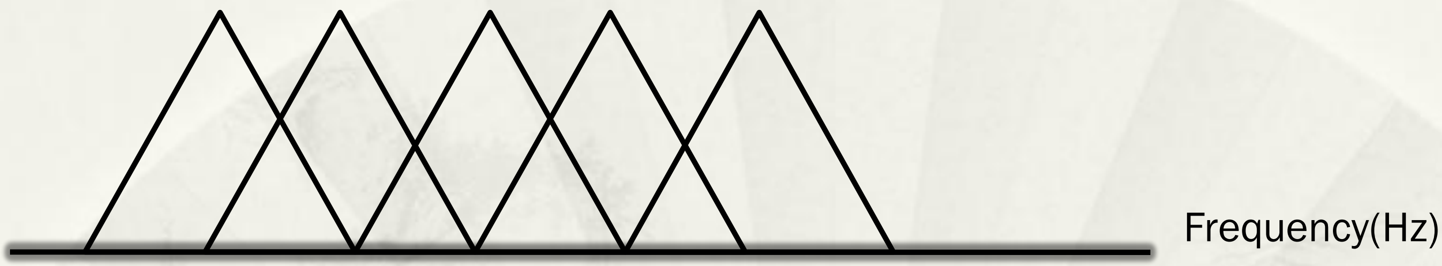
$\hat{f}_{min}$: the minimum frequency of the filter bank

N : the number of filters in the filter bank

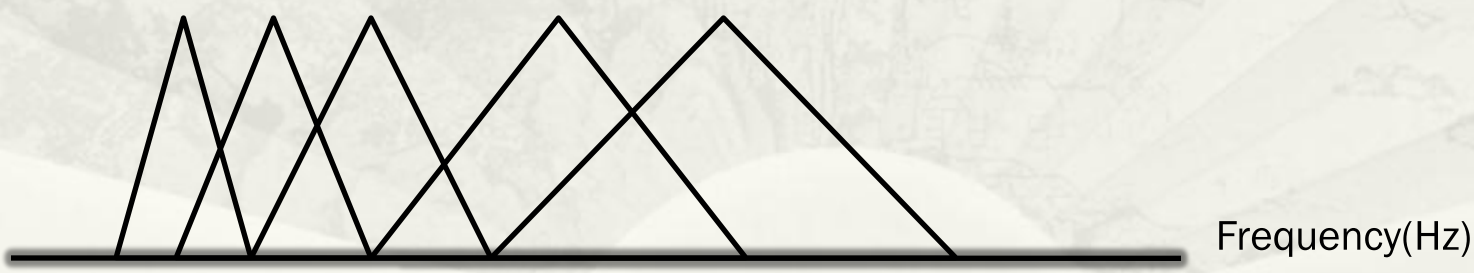
m : the percent overlap between adjacent filters bases $(0 \leq m \leq 1)$

Filter Widening Schemes(1)

- Mel-Frequency:



- Linear-Frequency:



Filter Widening Schemes(1)

- Since $\hat{f}_{max}$ and $\hat{f}_{min}$ are constant in this scheme, filter bank center $\hat{f}_n$ is a function of m :

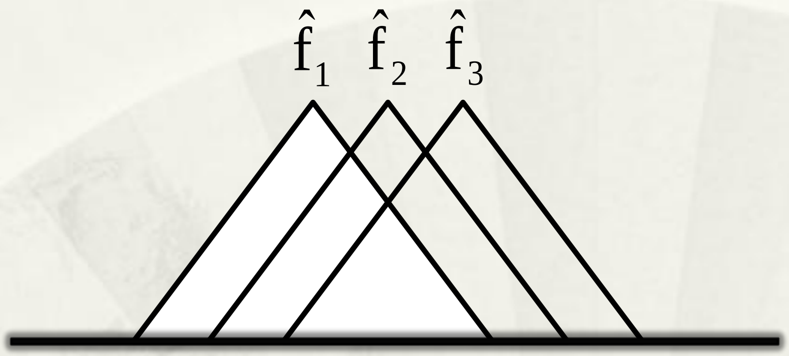
$$\hat{f}_n = \frac{L}{2} + (n-1)\frac{\hat{f}_{max} - \hat{f}_{min} - L}{N-1} + \hat{f}_{min}$$

$\hat{f}_n$: the center frequency of the n^{th} filter $(1 \le n \le N)$

- We refer to this algorithm as ***mfccVW*** since the filers are ***variable width*** according to the free parameter m .

Filter Widening Schemes(1)

◆ **Example:** $N=3$, $m=75\%$, $\hat{f}_{max}=1500$, $\hat{f}_{min}=0$



$$L=\frac{\hat{f}_{max}-\hat{f}_{min}}{N(1-m)+m}=\frac{1500-0}{3(1-0.75)+0.75}=1000$$

$$\frac{L}{2}=500\ ,\quad \frac{\hat{f}_{max}-\hat{f}_{min}-L}{N-1}=\frac{1500-0-1000}{3-1}=250$$

$$\hat{f}_1=500+250(1-1)=500$$

$$\hat{f}_2=500+250(2-1)=750$$

$$\hat{f}_3=500+250(3-1)=1000$$

Filter Widening Schemes(2)

- Equivalent rectangular bandwidth(ERB, in Hz) is the bandwidth of a rectangular filter centered at the center frequency of a critical band whose magnitude is maximum magnitude of the critical band and whose energy is the same as that of the critical band:

$$\mathrm{ERB} = \frac{\int \left|H(f)\right|^2 df}{\left|H(f_0)\right|^2}$$

- Using psychoacoustical measurements of ERB of human auditory filter with a center frequency of $\hat{f}_0$ (in kHz)

$$\mathrm{ERB} = 6.23 f_0^2 + 93.39 f_0 + 28.52$$

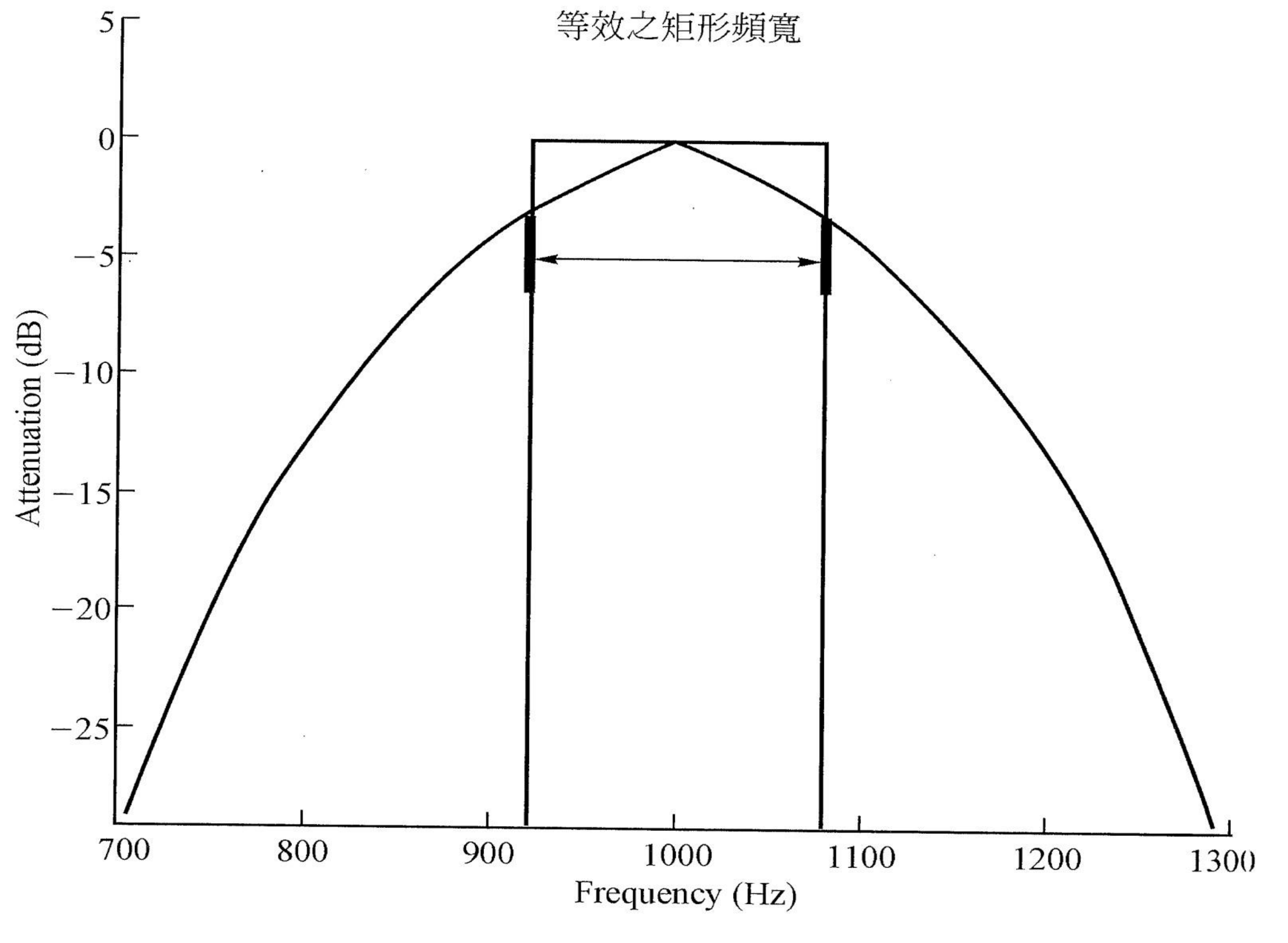

等效之矩形頻寬
5
0
−5
−10
−15
−20
−25
Attenuation (dB)
700
800
900
1000
1100
1200
1300
Frequency (Hz)

Fig. ERB

Filter Widening Schemes(2)

- The center frequencies used by the traditional mfcc are kept constant, and the ERB for each center frequency is calculated.
- The mel-frequency warping function between linear frequency f and mel-frequency $\hat{f}$ is :

$$\hat{f}=2595\log_{10}(1+\frac{f}{700})$$

and

$$\hat{f}_0=\frac{1}{2}(\hat{f}_H+\hat{f}_L)\ ;\ \text{ERB}=\frac{1}{3}(f_H-f_L)$$

are solved for f_H and f_L when f_0 is given and ERB is calculated.

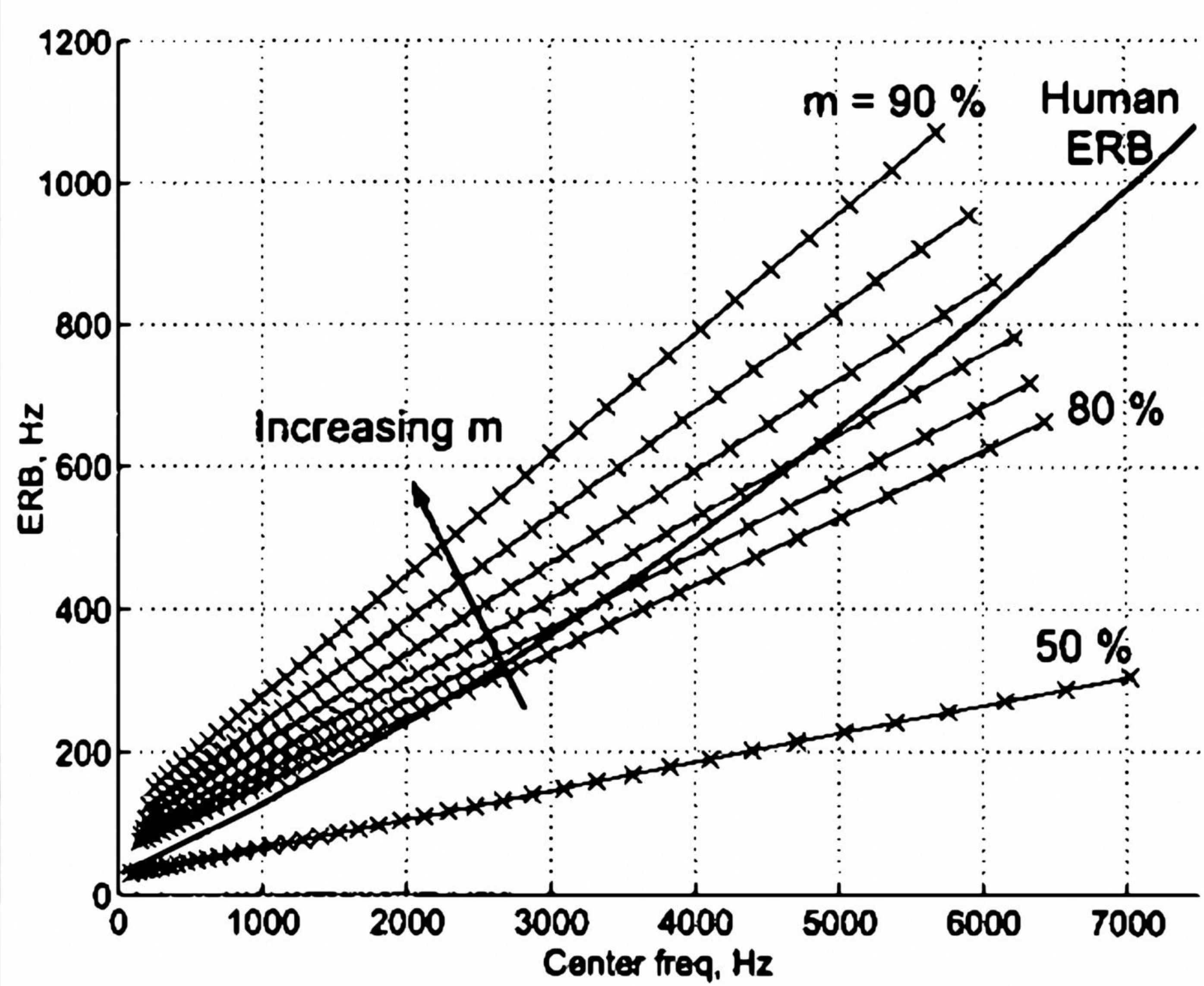


Fig1. ERB vs Frequency for mfcc (m=50%) and mfccVW (m=80%,82%,...,90%) as well as for the human auditory system . Each x corresponds to a filter center frequency.

Experiment And Conclusion

- To characterize our modified filter banks, we perform two experiments on vowel extracted from the TIMIT database.

  The vocabulary consists of 10 vowels (/IY/,/IH/,/EH/,/AE/,/AA/,/UH/,/UW/,/AH/,/ER/) extracted from read sentences according to the phonetic labels provided by the corpus(~50,000 phonemes in all).

- The first experiment using the Fisher discriminant (J-measure)
- The second experiment using a Bayes classifier.

Experiment (1)

- The first experiment measures the Fisher discriminant (J-measure) for the 10-class problem. This measure compares variance between classes to variance within each class.
- Larger J-measures denote greater separation between classes.

$$J=\mathrm{trace}(S_W^{-1}S_B)$$

$$S_W=\sum_{k=1}^{c}S_k=\sum_{k=1}^{c}N_k\sum k \qquad S_B=\sum_{k=1}^{c}N_k(m_k-m_0)(m_k-m_0)^T$$

S_w : the within-class scatter

S_B : the between-class scatter

m_0: the mean vector of the entire data set

N_k: the number of samples in the k^{th} class(c classes total)

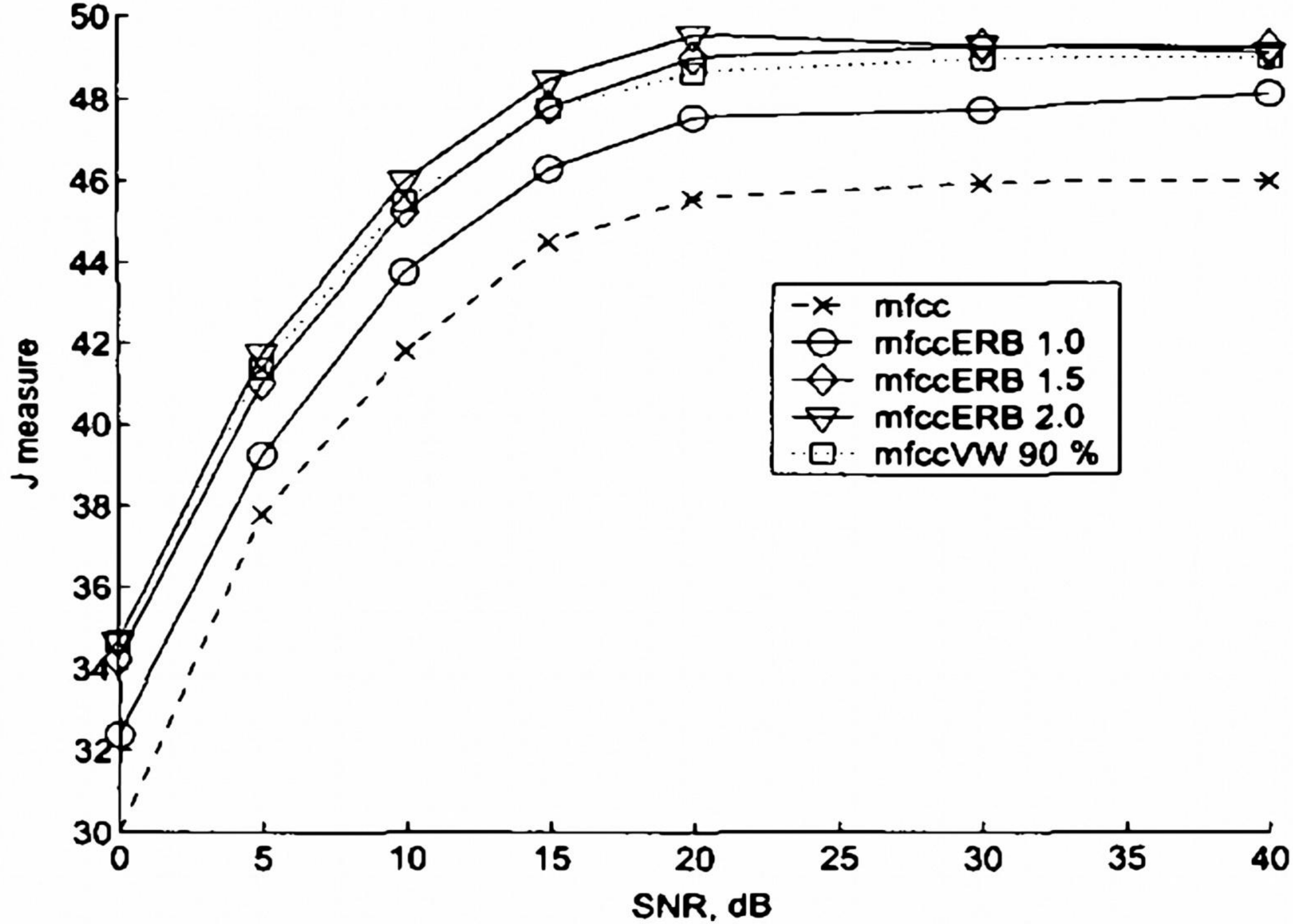


Fig. 2. J-measure vs SNR for mfcc, mfccVW, and mfccERB (inflation factors 1.0, 1.5, and 2.0).

Experiment (2)

- The second experiment using a Bayes classifier, each class is divided into test and train data(80% train).
- Classification is determined by maximizing the discriminant function:

$$g_k = x^T W_k x + w_k{}^T x + \omega_{k0}$$

where

$$W_k = -\frac{1}{2}\sum_k^{-1}$$

$$w = \sum_k^{-1} m_k$$

$$\omega_{ko} = -\frac{1}{2} m_k^T \sum_k^{-1} m_k - \frac{1}{2}\log\left|\sum k\right| + \log P_k$$

for covariance $\sum k$ and mean m_k of the k^{th} class with a priori probability P_k . x is the vector of cepstral coefficients for the test data and is classified as $\arg(\max_k g_k)$.

The mfccVW results

(1) Performance at all SNRs is nearly maximized for m near ***90%***.

(2) The dramatic change in recognition results for m ***> 90%.***

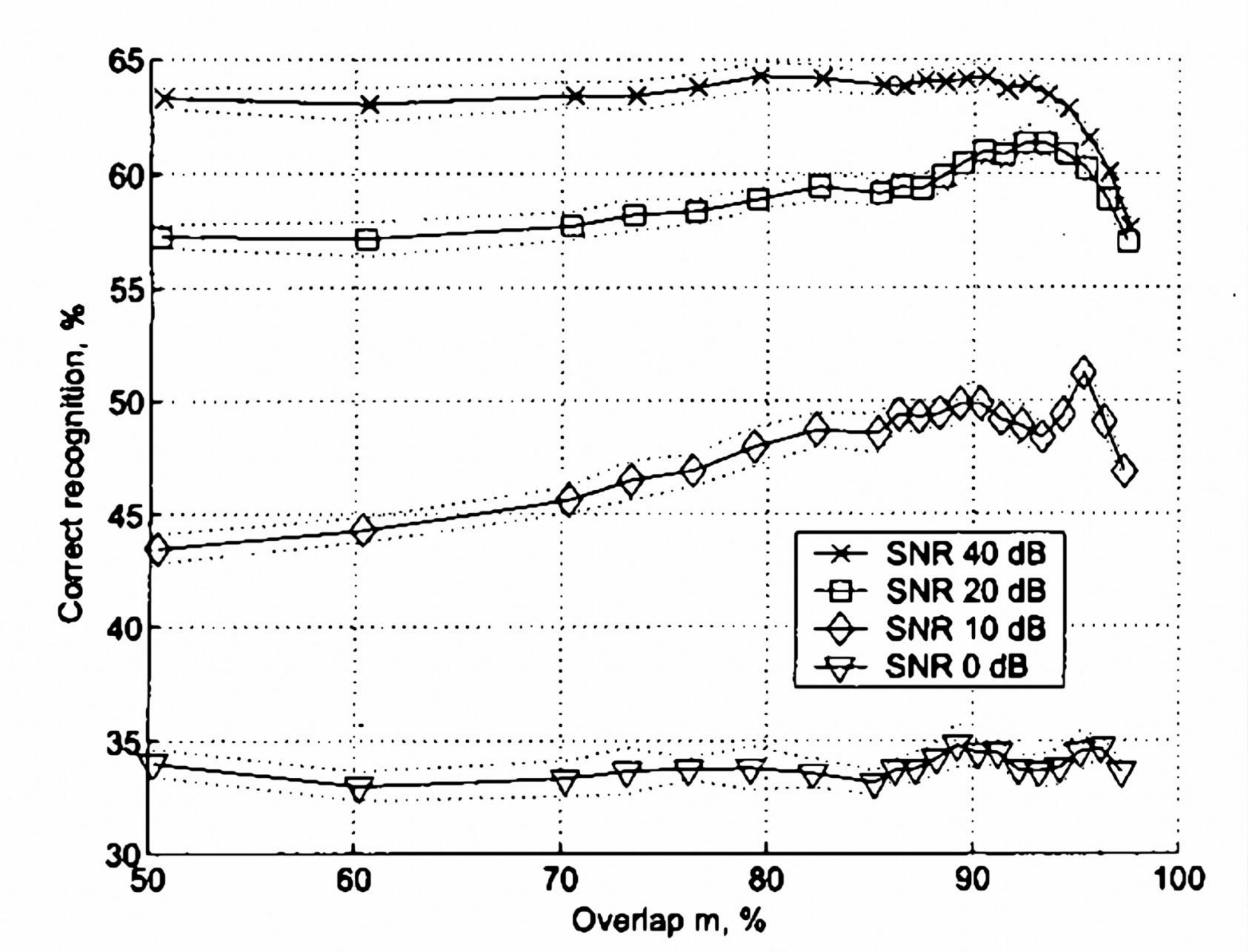


Fig.3 (*a*)mfcc and mfccVW for various SNRs vs Overlap percentage m

The mfccERB results

(1) Recognition is nearly the same between 30-40 dB SNR while the other filter banks increased in performance over the same range .

(2) The results using the largest mfccERB scale factor are highest for moderate noise (5-30 dB).

(3) Below 5 dB SNR

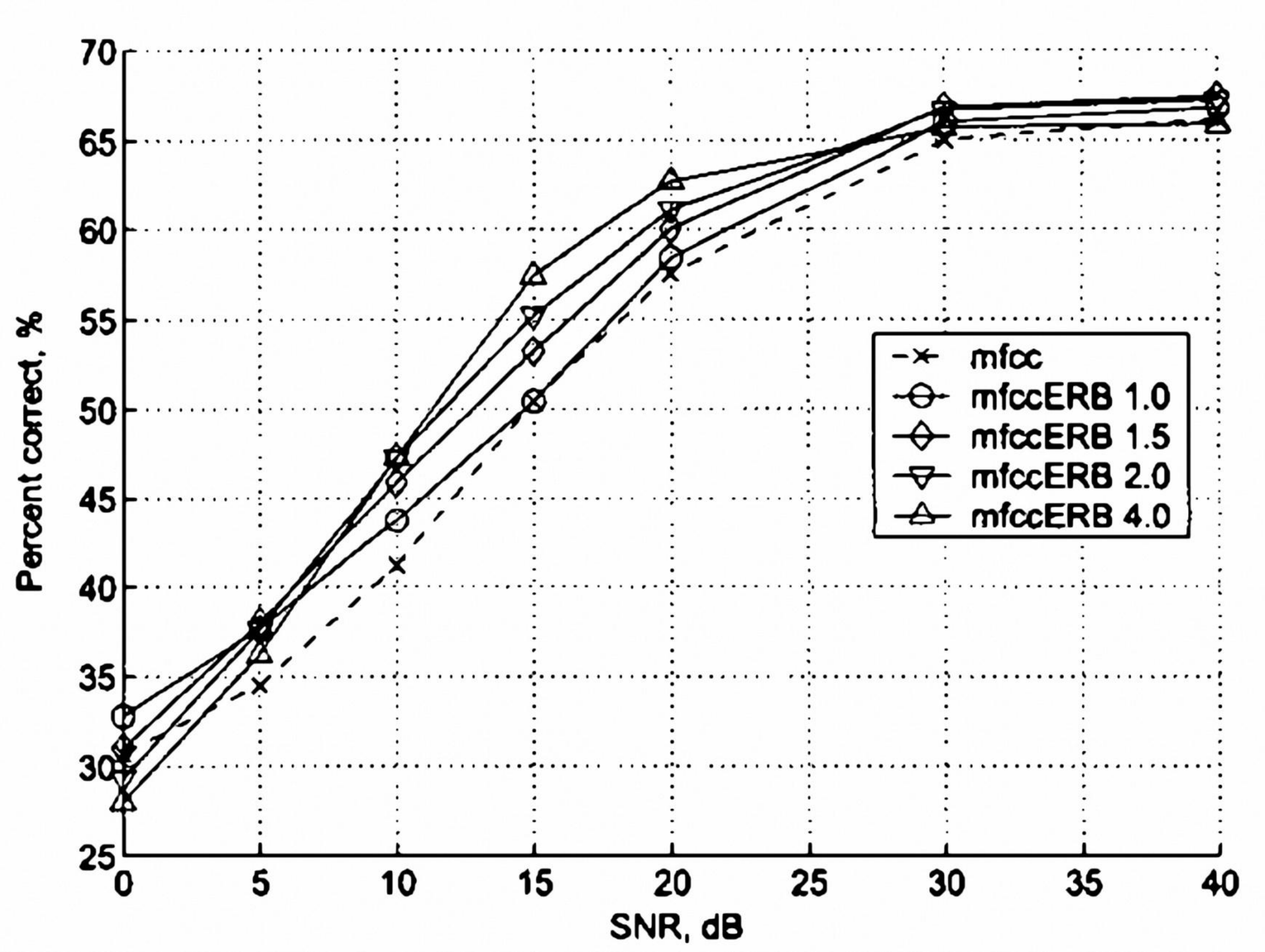


Fig.3 (*b*) mfcc and mfccERB with ERB scale factors 1.0 , 1.5 , 2.0 and 4.0