Incorporation Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm

Author : Weizhong Zhu
Douglas O'Shaughnessy
Professor: 陳嘉平
Reporter: 吳國豪

Outline

- Introduction
- Frequency masking model
- Experiment
- Conclusions

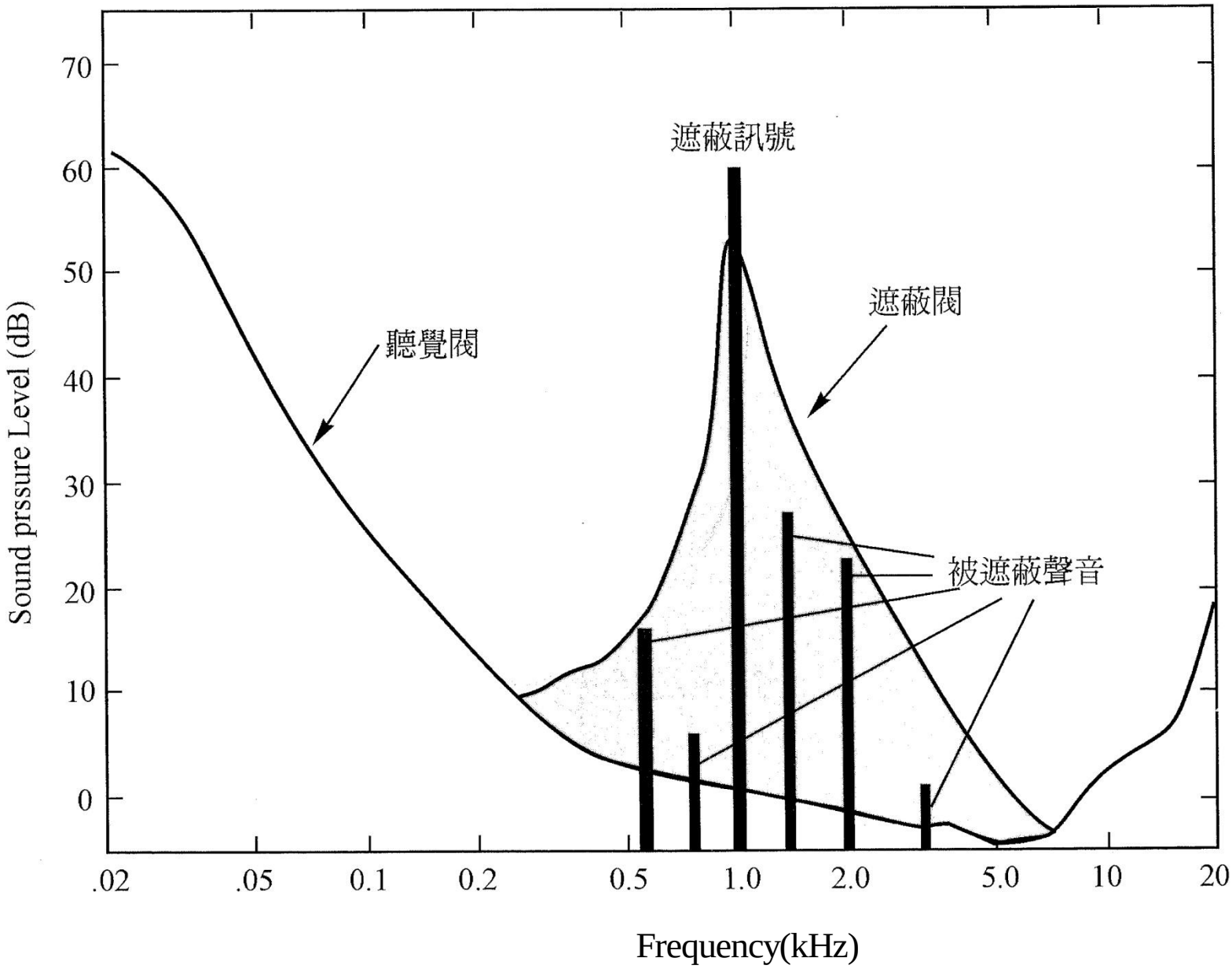
Introduction

- Accuracy of speech recognition degrades rapidly when speech is distorted by noise.

- Methods to overcome the effects of noise must be applied in order to achieve good recognition accuracy in real speech recognition applications where various types of noises may exist.

- When unknown noise is involved, training-test mismatch always occurs, so that the performance of ASR degrades significantly in noisy conditions. The human auditory system has some mechanism to reduce the effect of noise .

- An important aspect of human audition is the phenomenon of masking.


70
60
50
40
30
20
10
0
Sound prssure Level (dB)
遮蔽訊號
聽覺閥
遮蔽閥
被遮蔽聲音
.02
.05
0.1
0.2
0.5
1.0
2.0
5.0
10
20
Frequency(kHz)

Introduction

- The aim of this paper is to introduce frequency masking filtering in a standard Mel Frequency Cepstral Coefficients(MFCCs) feature extraction algorithm.
- We want to answer the following two questions:

  (1) Comparing with standard MFCC, how good are the performances in the sense of relative improvement if we introduce the frequency masking filtering with different thresholds?

  (2) Can frequency masking filtering be used to combine with another proved noise robustness technique, such as cepstral mean normalization?

Frequency masking model

- The essence of the masking is to enhance a dominant signal component and to suppress the noise components.
- Usually a speech parameter is extracted from its power spectrum.
- To make the frequency masking model simple, the spectral shape of a frequency masking threshold is modeled as a triangle.
- The model has two parameters, α for the masking threshold of lower-frequency and β for the masking threshold of higher-frequency.

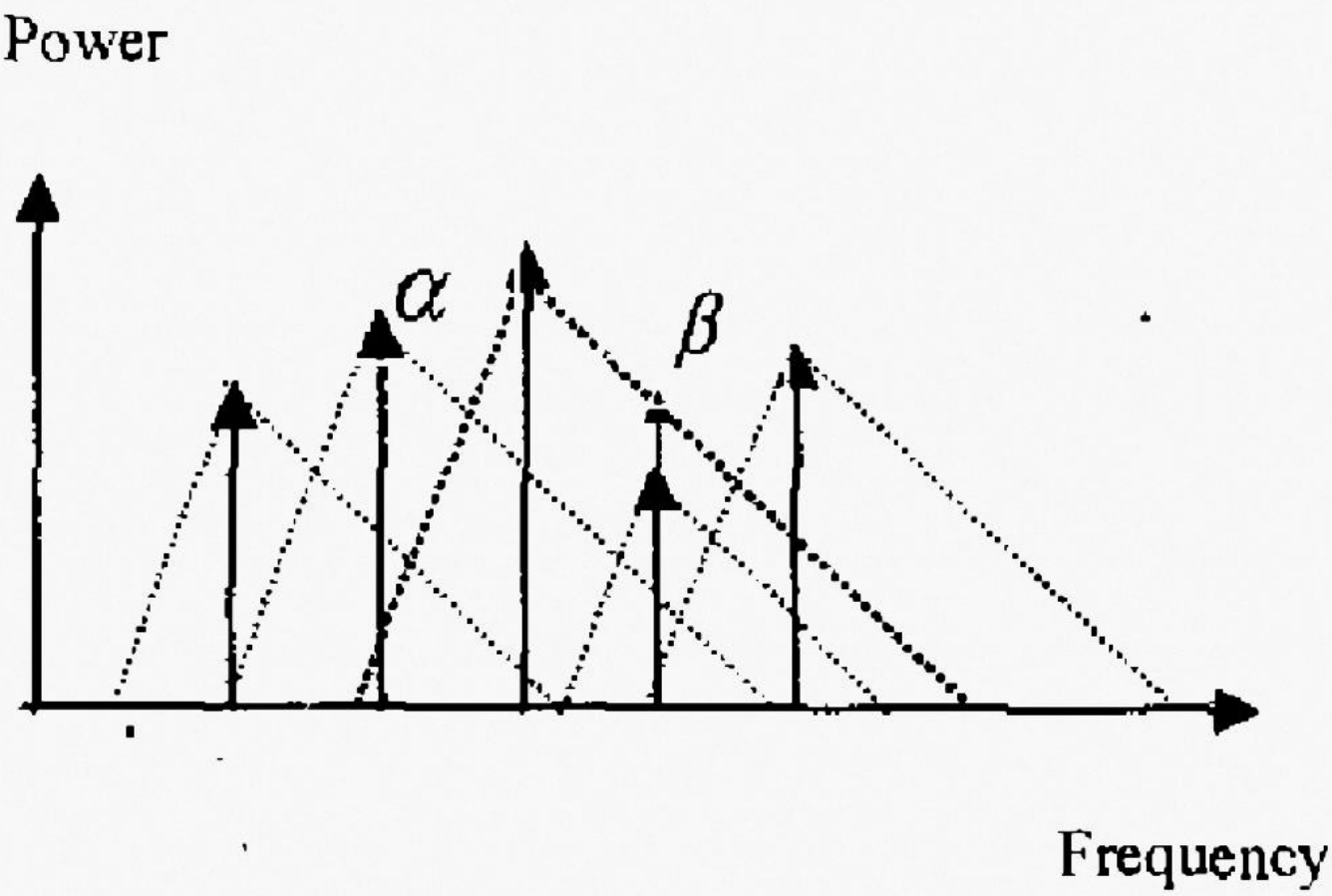


Figure 1: Schematic representation of Frequency masking.

Frequency masking model

$$y'_{i-1} = \alpha\, y_i$$

$$y_{i-1} = y'_{i-1} \quad if\ y'_{i-1} > x_{i-1} \qquad (1)$$

$$y_{i-1} = x_{i-1} \quad if\ y'_{i-1} \le x_{i-1}$$

where x_i is the original power spectrum at frequency index i and y_i is the output of the filtered spectrum, α is the masking threshold of lower-frequency. Equation 1 is executed from the higher frequency index to the lower one with initialization of $y_N = x_N$.

$$y'_n = \beta\, y_{n-1}$$

$$y_n = y'_n \quad if\ y'_n > x_n \qquad (2)$$

$$y_n = x_n \quad if\ y'_n \le x_n$$

where β is the masking threshold of higher-frequency. Equation 2 is executed in the opposite direction with initialization of $y_0 = x_0$.

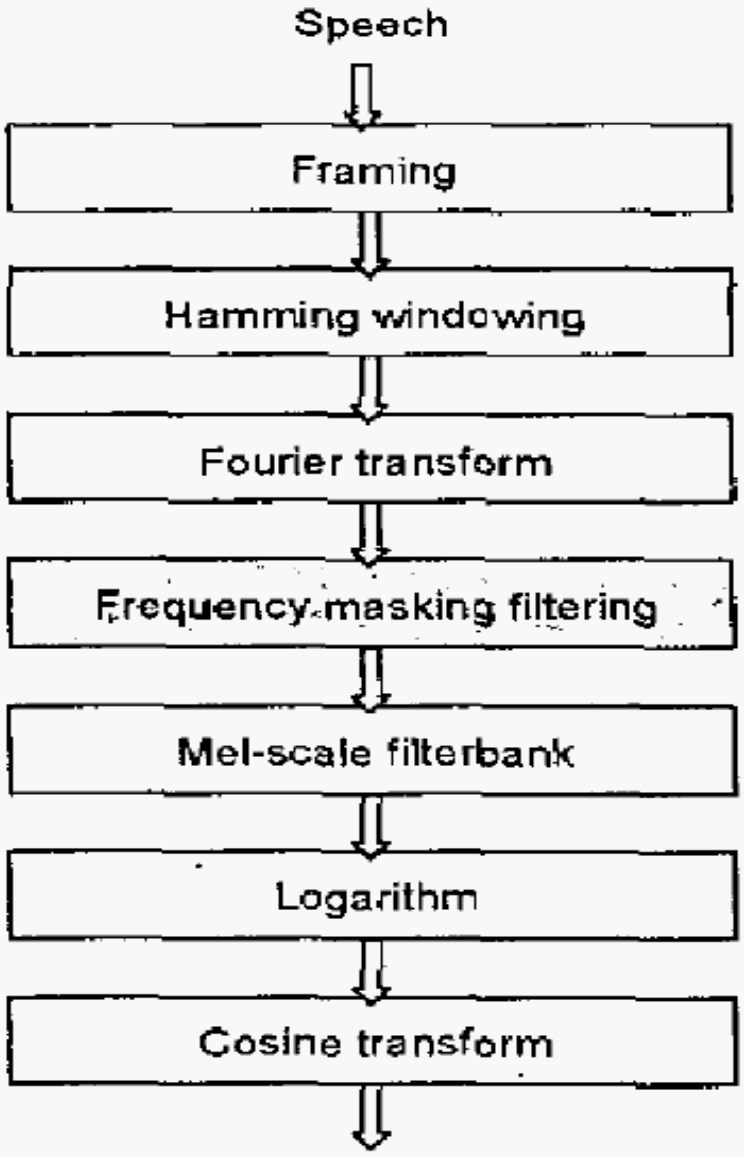


Figure 2: Flowchart of MFCC with frequency masking.

Experiment

- The proposed method was evaluated on the Aurora 2 database. All recognition tests were conducted using the HTK recognition toolkit with the setting defined for evaluation. The task is speaker-independent recognition of digit sequences. The distortions are artificially added to the clean TIdigits database . Selections of 8 different real-world noises have been added to the speech over a range of signal-to-noise ratios(SNRs: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, clean – no noise added).

Experiment

- There are three tests from the Aurora 2 database to evaluate the performance of all considered techniques. Data in Test A are added to by noises of Subway, Babble, Car and Exhibition. Data in Test B are added to noises of Restaurant, Street. Airport and Station. In Test C, besides the additive noise, channel distortion is also included.
- The results in this section are defined in terms of relative improvement *(RI.). According to the Aurora 2 protocol, it is* calculated as

$$R.I = \frac{NewScore - Baseline}{100 - Baseline} \times 100\%$$

where NewScore, Baseline are recognition accuracies for each test using proposed and reference algorithms, respectively. The mean recognition accuracy for each test set is obtained by taking the average of the recognition accuracies measured in 20. 15. 10. 5 and 0 dB SNR. The overall accuracy is calculated as 0.4*Set A + 0.4*Set B + 0.2*Set C.

Experiment 1

- In experiment 1, we explore how good the performances are in the sense of relative improvement if we introduce the frequency masking filtering with different thresholds. What are the optimized parameters?
- We know that, basically, both the masking thresholds of α and β can be values between 0 and 1. Psychoacoustic experiments showed that lower frequency tends to mask higher frequency. Our experiments also confirmed that in the condition of $\alpha \leq \beta$, it gives better results than that of the opposite. Table 1 shows the overall relative improvements with various thresholds.
- We found that except for β = 0.9, all the combinations have positive effect.
- The relative improvement is raised as the masking threshold of higher frequency varies from 0.2 to 0.8. it achieves a 4.47% highest relative improvement when α = 0.5 and β = 0.8.

Table 1: Relative improvements (%) in different combinations of thresholds.

$\beta \backslash \alpha$	0.2	0.3	0.4	0.5	0.6	0.7
0.2	0.16					
0.3	0.41	0.59				
0.4	1.01	1.03	1.27			
0.5	1.69	1.85	2.01	2.12		
0.6	2.83	2.89	3.00	2.88	2.48	
0.7	3.83	3.83	4.04	4.00	3.54	1.62
0.8	4.13	4.19	4.36	4.47	4.11	2.70
0.9	-4.64	-4.59	-4.44	-4.05	-3.77	-3.98

Experiment 2

- In order to be consistent with results of psychoacoustic experiments, in which the slope of triangle masking threshold should be in a logarithmic frequency scale. We use a linear interpolation method to compensate it.

- Since MFCCs are derived from 23 Mel-scale triangular filters, we define two arrays of α_i and β_i for each frequency index i. α_i is calculated as a linear interpolation of 0.3 and 0.5, while β_i is calculated as a linear interpolation of 0.6 and 0.8 along with 23 Mel-scale triangular filters.

- Table 2 shows performance comparison between fixed thresholds (α = 0.5 and β = 0.8) and their linear interpolations for average relative improvements at different SNR levels. This method not only increases overall relative improvement, but also reduces the negative effect at higher SNRs

Table 2: Performance comparison between fixed thresholds and their linear interpolations for average relative improvements (%) at different SNR levels.

	clean	20 dB	15 dB	10 dB	5 dB	0 dB
Fixed	-10.39	-6.36	4.66	7.76	'5.60	3.07
L.I.	-1.23	1.36	10.64	9.47	6.09	2.54

Table 3: Relative improvements of proposed algorithm over standard MFCC under the evaluation paradigm specified by Aurora 2.

Clean Training - Relative Performance														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	0.00%	-9.00%	5.77%	3.75%	0.13%	0.00%	-9.00%	5.77%	3.75%	0.13%	-10.47%	-2.91%	-6.69%	-1.26%
20 dB	1.02%	16.85%	1.93%	-19.67%	0.03%	11.09%	8.45%	16.88%	10.61%	11.76%	-20.64%	-12.94%	-16.79%	-1.36%
15 dB	12.29%	11.85%	28.71%	8.92%	15.44%	8.38%	12.03%	13.75%	20.73%	13.72%	-8.84%	-1.44%	-5.14%	1.54%
10 dB	17.06%	5.68%	18.01%	13.68%	13.61%	7.14%	13.68%	8.28%	15.16%	11.07%	-4.60%	0.59%	-2.01%	9.47%
5 dB	11.06%	2.98%	8.24%	9.01%	7.82%	3.46%	7.77%	4.19%	6.73%	5.54%	2.15%	5.36%	3.76%	6.09%
0 dB	4.19%	1.50%	1.89%	3.36%	2.73%	2.00%	2.42%	2.25%	1.64%	2.08%	3.22%	2.94%	3.08%	2.54%
-5dB	1.24%	0.53%	0.50%	0.64%	0.73%	0.98%	0.57%	0.03%	0.37%	0.49%	1.33%	0.32%	0.82%	0.65%
Average	8.42%	4.46%	8.07%	6.39%	6.59%	4.43%	6.77%	5.74%	7.33%	6.01%	-0.16%	2.57%	1.20%	5.42%

A					
	Subway	Babble	Car	Exhibition	Average
Clean	0.00%	-9.00%	5.77%	3.75%	0.13%
20dB	1.02%	16.85%	1.93%	-19.67%	0.03%
15dB	12.29%	11.85%	28.71%	8.92%	15.44%
10dB	17.16%	5.68%	18.01%	13.68%	13.61%
5dB	11.06%	2.98%	8.24%	9.01%	7.82%
0dB	4.19%	1.50%	1.89%	3.36%	2.73%
-5dB	1.24%	0.53%	0.50%	0.64%	0.73%
Average	8.42%	4.46%	8.07%	6.39%	6.59%

B					
	Restaurant	Street	Airport	Station	Average
Clean	0.00%	-9.00%	5.77%	3.75%	0.13%
20dB	11.09%	8.45%	16.88%	10.61%	11.76%
15dB	8.38%	12.03%	13.75%	20.73%	13.72%
10dB	7.14%	13.68%	8.28%	15.16%	11.07%
5dB	3.46%	7.77%	4.19%	6.73%	5.54%
0dB	2.00%	2.42%	2.25%	1.64%	2.08%
-5dB	0.98%	0.57%	0.03%	0.37%	0.49%
Average	4.43%	6.77%	5.74%	7.33%	6.01%

C			
	Subway	Street	Average
Clean	-10.47%	-2.91%	-6.69%
20dB	-20.64%	-12.94%	-16.79%
15dB	-8.84%	-1.44%	-5.14%
10dB	-4.60%	0.59%	-2.01%
5dB	2.15%	5.36%	3.76%
0dB	3.22%	2.94%	3.08%
-5dB	1.33%	0.37%	0.82%
Average	-0.16%	2.57%	1.20%

Average	
Clean	-1.23%
20dB	1.36%
15dB	10.64%
10dB	9.47%
5dB	6.09%
0dB	2.54%
-5dB	0.65%
Average	5.42%

Experiment 3

- We know that some robust feature extraction techniques could get a better result. Can our proposed algorithm combine with these techniques to get an even better result? Here in experiment 3. we answer the question.

- Researchers have reported a significant improvement using Cepstral Mean Normalization (CMN). We tested the CMN method. Each MFCC vector is reduced by its utterance mean. If we use CMN only, we can get an overall 19.30% relative improvement. When we combine CMN with our frequency masking filtering algorithm, the recognition performance gain can be further improved to 22.74%.

Table 4: Summary of relative improvements of techniques with respect to the performance of the standard MFCC.

	Set A	Set B	Set C	Overall
FM1	5.75%	4.10%	2.58%	4.47%
FM2	6.59%	6.01%	1.20%	5.42%
CMN	12.51%	34.05%	-3.73%	19.30%
CMN+FM2	16.54%	36.61%	0.64%	22.74%

Conclusions

- The proposed method is effective to improve the performance of speech recognition for eight different noise conditions at various SNR levels.

- The proposed frequency masking filtering algorithm can easily be embedded in a standard front-end MFCC calculation program with a very small extra computation load.