Feature Extraction for Robust Speech Recognition using a Power-Law Nonlinearity and Power-Bias Subtraction

Author : Richard M. Stern ,
Chanwoo Kim
Professor : 陳嘉平
Reporter : 許峰閤

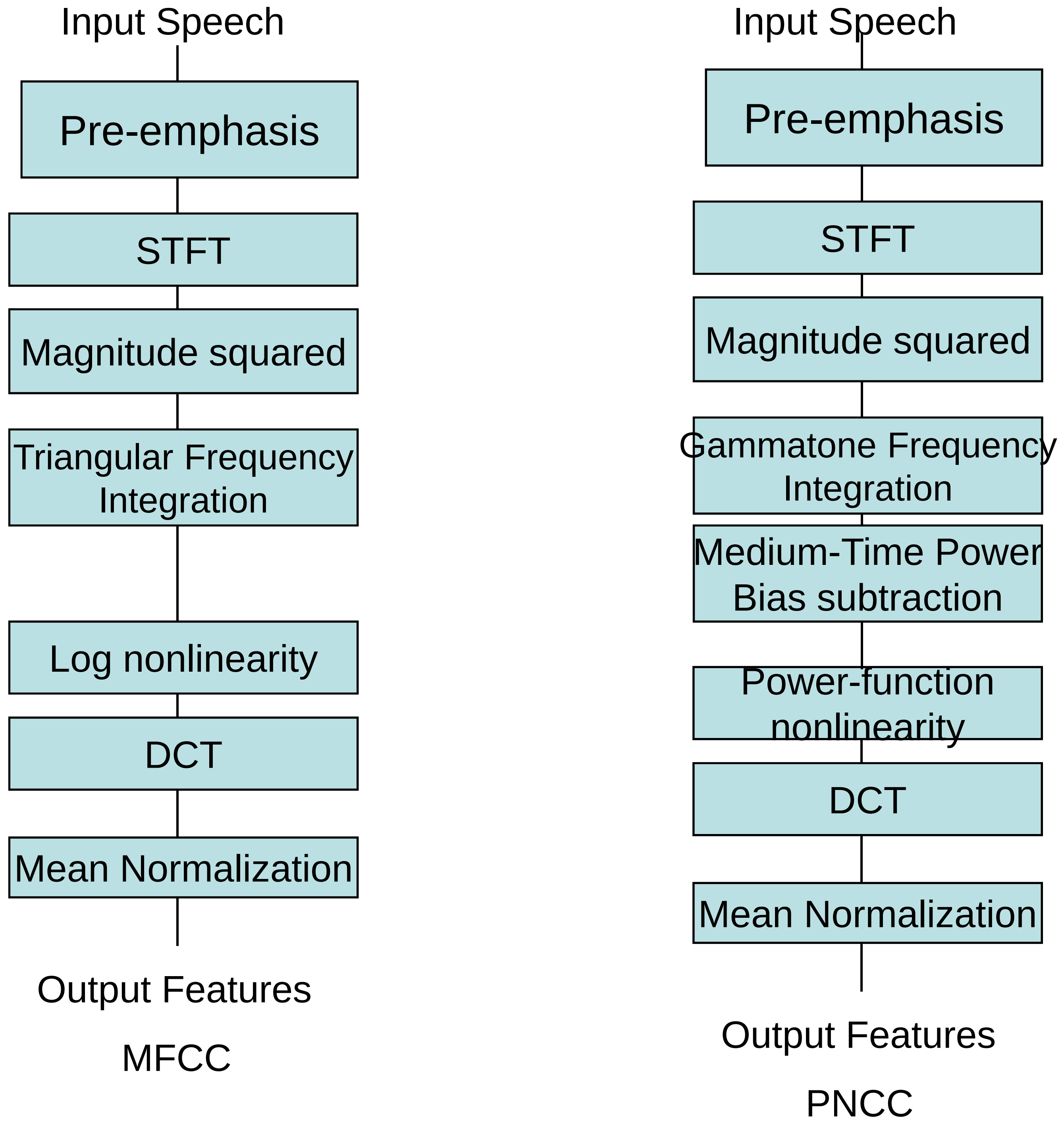
Outline

- Introduction
- Power-Law Function
- Mdium-duration power bias removal

Introduction

- Power-Normalized Cepstral Coefficient (PNCC) is a new feature extraction algorithm. It’s provides great improvements in recognition accuracy compared to MFCC and PLP.

- Using Power-bias subtraction to increase the speech accuracy.


Input Speech
Pre-emphasis
STFT
Magnitude squared
Triangular Frequency Integration
Log nonlinearity
DCT
Mean Normalization
Output Features
MFCC
Input Speech
Pre-emphasis
STFT
Magnitude squared
Gammatone Frequency Integration
Medium-Time Power Bias subtraction
Power-function nonlinearity
DCT
Mean Normalization
Output Features
PNCC

Power-Law Function

- Because the logarithmic nonlinearity used in MFCC does not exhibit threshold behavior.

- The logarithmic would produce a large output change even if the changes in input are small

$$y = x^{a0}$$

Medium-duration power bias removal

- The medium-duration power bias removal provides further decrease in WER.

- It’s consist of estimating B(i) and then computing the system output that would be obtained after removed .

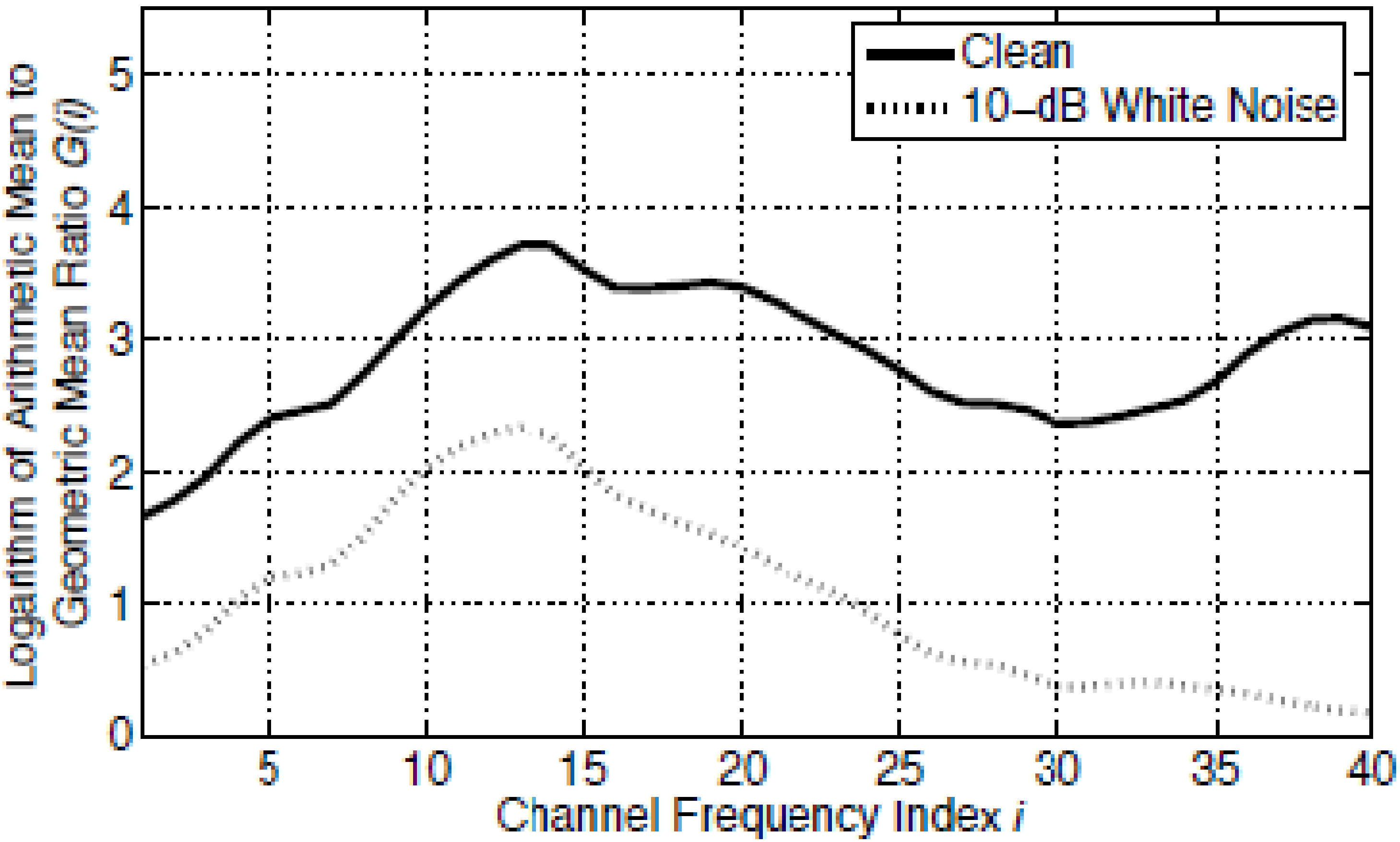
- Estimate the medium-duration power of speech signal Q(i,j) by computing the running average of P(i,j) .

$$Q(i,j)=\frac{1}{2M+1}\sum_{j'=j-M}^{j+M}P(i,j')$$

i represents channel index and j is frame index . M=3 is optimal

- We find it convenient to use the ratio of arithmetic mean to geometric mean
 (AM to GM ratio) to estimate the degree of speech corruption .

$$G(i) = \log[\sum_{j=0}^{J-1} \max(Q(i,j),\$)] - \frac{1}{J}\sum_{j=0}^{J-1} \log[\max(Q(i,j),\$)]$$



Clean
10–dB White Noise
Logarithm of Arithmetic Mean to Geometric Mean Ratio G(i)
Channel Frequency Index i
0
1
2
3
4
5
5
10
15
20
25
30
35
40

Removing the power bias

- Power bias is B(i)
- The normalized power $Q'(i, j \mid B(i))$ is given by following equation:

$$Q'(i, j \mid B(i)) = \max(Q(i, j) - B(i), d_0 Q(i, j))$$

- Define the parameter $G'(i \mid B(i))$

$$G'(i \mid B(i)) = \log[\sum_{j=0}^{J-1} \max(Q'(i, j \mid B(i)), C_f(i))]$$
$$-\frac{1}{J}\sum_{j=0}^{J-1} \log[\max(Q'(i, j \mid B(i)), C_f(i))]$$

$$C_f(i) = d_1(\frac{1}{J}\sum_{j'=0}^{J-1} Q(i, j'))$$

- We noted that G(i) statistic is smaller for corrupt speech than it is for clean speech.

- We can define the power bias $B^*(i)$ as the smallest power makes the G(i) the same as that of clean speech.

$$B^*(i) = \min\{B(i) \mid G'(i \mid B(i)) \geq G_{cl}(i)\}$$

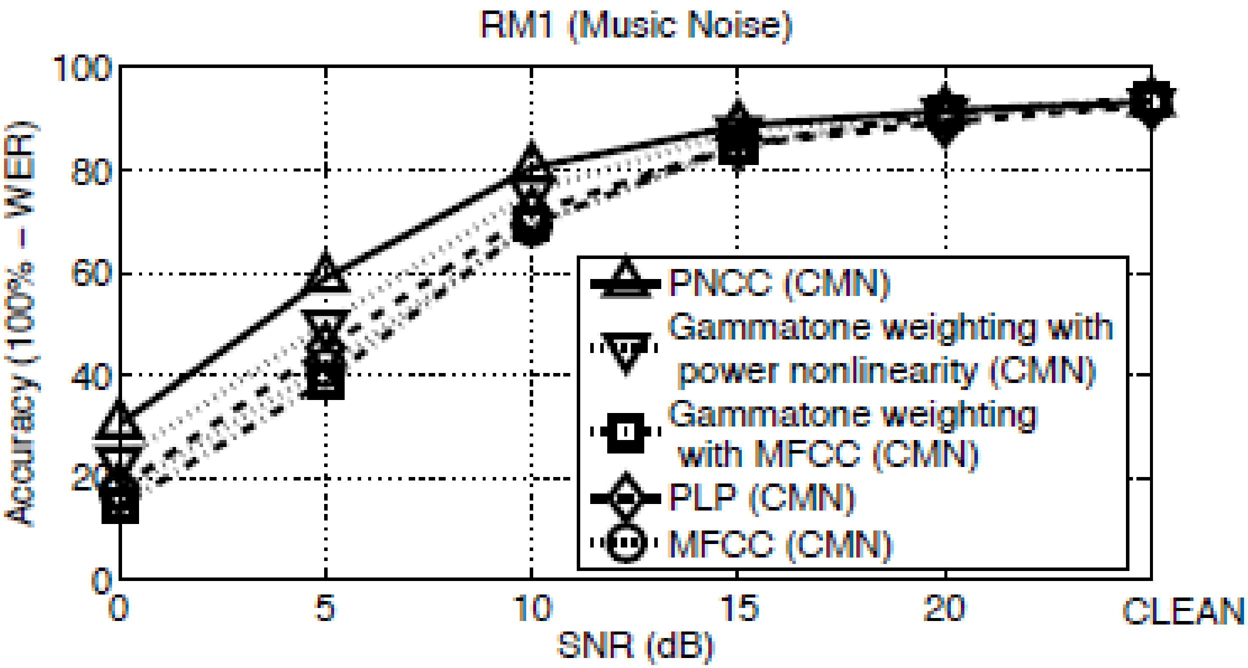
- Using this procedure for each channel , we can obtain $Q'(i,j\,|\,B^{*}(i))$
- For each time-frequency bin represented by (i,j) , the power normalization gain is given by:

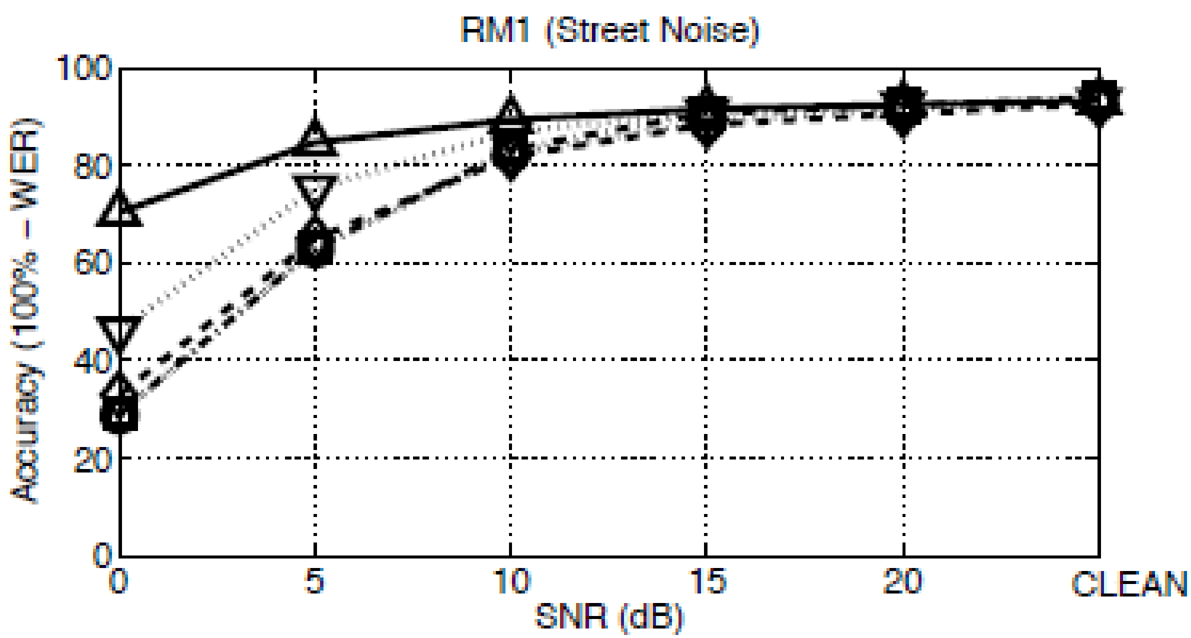
$$w(i,j)=\frac{Q'(i,j\,|\,B^{*}(i))}{Q(i,j)}$$

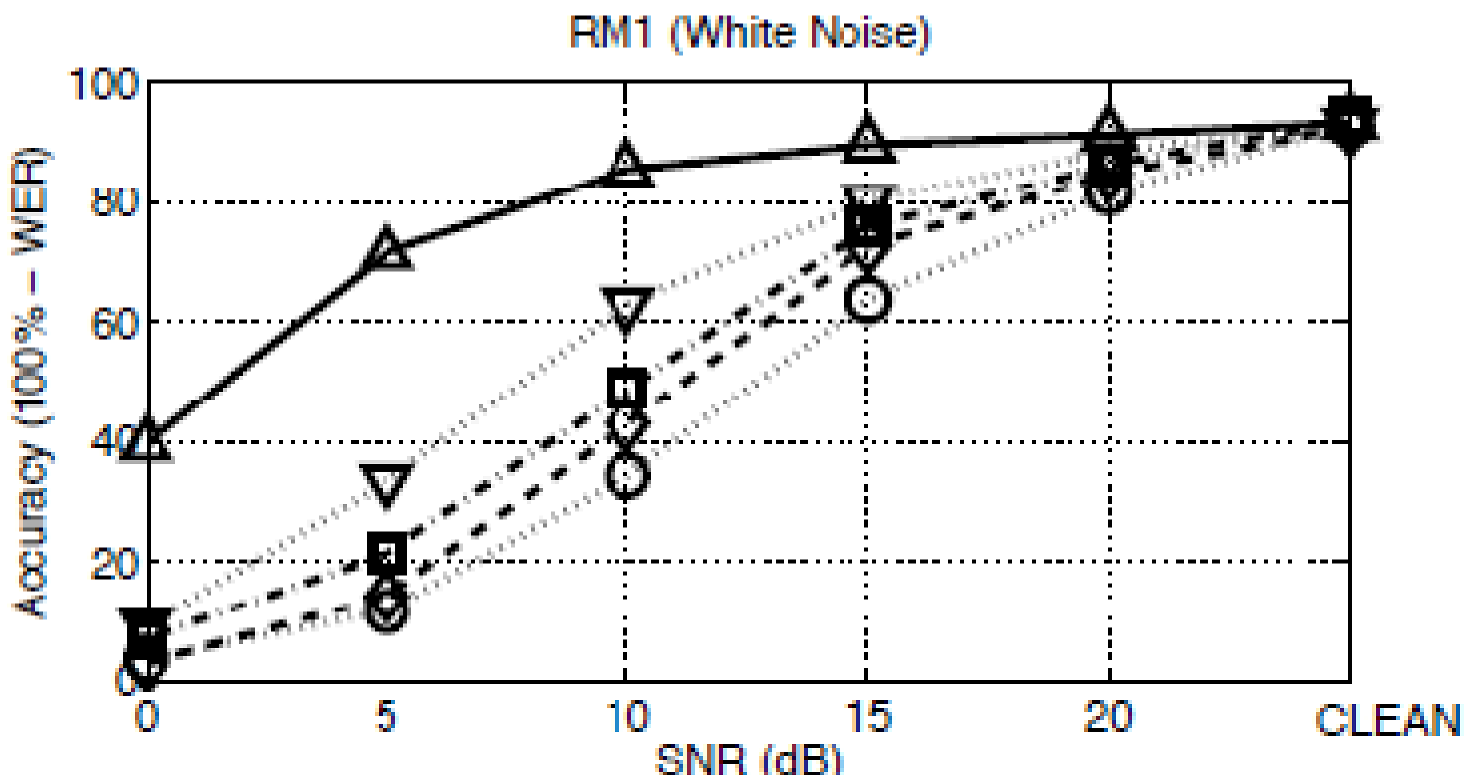
- For smoothing purposes , we average across channels from the i-Nth channel to i+Nth , thus the final power $P'(i,j)$ is given by the following equation:

$$P'(i,j) = \left(\frac{1}{2N+1}\sum_{i'=\max(i-N,1)}^{\min(i+N,C)} w(i',j)\right)P(i,j)$$

- C is total number of channels . N=5 and total number of 40 gammatone channels


RM1 (Music Noise)
Accuracy (100% − WER)
SNR (dB)
0
5
10
15
20
CLEAN
0
20
40
60
80
100
PNCC (CMN)
Gammatone weighting with power nonlinearity (CMN)
Gammatone weighting with MFCC (CMN)
PLP (CMN)
MFCC (CMN)


RM1 (Street Noise)
Accuracy (100% − WER)
SNR (dB)
0
5
10
15
20
CLEAN
0
20
40
60
80
100


RM1 (White Noise)
Accuracy (100% – WER)
100
80
60
40
20
0
0
5
10
15
20
CLEAN
SNR (dB)