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

*Chanwoo Kim*[2], *Richard M. Stern*[1,2]

[1]Department of Electrical and Computer Engineering
and [2]Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA 15213 USA
{chanwook, rms}@cs.cmu.edu

## Abstract

This paper presents a new feature extraction algorithm called Power-Normalized Cepstral Coefficients (PNCC) that is based on auditory processing. Major new features of PNCC processing include the use of a power-law nonlinearity that replaces the traditional log nonlinearity used for MFCC coefficients, and a novel algorithm that suppresses background excitation by estimating SNR based on the ratio of the arithmetic to geometric mean power, and subtracts the inferred background power. Experimental results demonstrate that the PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for various types of additive noise. The computational cost of PNCC is only slightly greater than that of conventional MFCC processing.

**Index Terms**: Robust speech recognition, physiological modeling, rate-level curve, power function, ratio of arithmetic mean to geometric mean, power distribution normalization

## 1. Introduction

Even though many speech recognition systems have obtained satisfactory performance in clean environments, recognition accuracy significantly degrades if the test environment is different from the training environment. These environmental differences might be due to additive noise, channel distortion, acoustical differences between different speakers, and so on. Many algorithms have been developed to enhance the environmental robustness of speech recognition systems. Figure 1 compares the structure of conventional MFCC processing, PLP processing [1], and the new approach described in this paper, which will be called Power-Normalized Cepstral Coefficients (PNCC). As can be seen from Fig. 1, the major innovations in this algorithm are the use of a well-motivated power function that replaces the log function, and the use of a novel approach to the blind removal of background excitation based on medium-duration power estimation. This normalization makes use of the ratio of the arithmetic mean to the geometric mean, which has proved to be a useful measure in determining the extent to which speech is corrupted by noise [2]. In addition, PNCC uses frequency weighting based on the gammatone filter shape [3] rather than the triangular frequency weighting or the trapezoidal frequency weighting associated with the MFCC and PLP computation, respectively. A pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is applied first. The STFT analysis is performed using Hamming windows of duration 25.6 ms, with 10 ms between frames for a sampling frequency of 16 kHz, 40 gammatone channels. After passing through the gammatone channel, the power is normalized using peak power (*i.e.* the $95th$ percentile of short-time power).

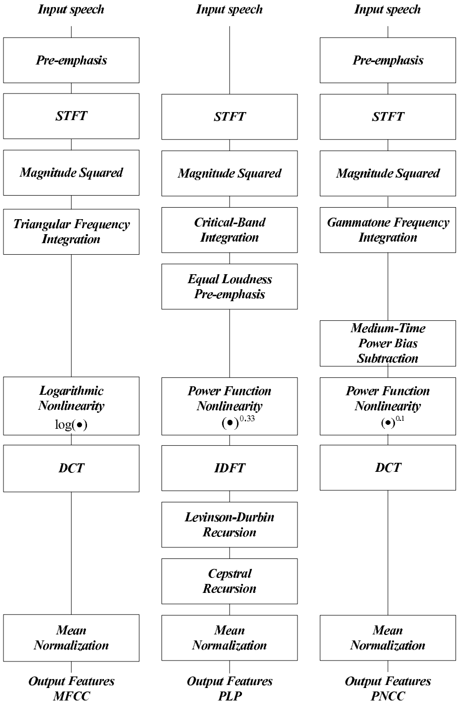

Input speech
Input speech
Input speech
Pre-emphasis
Pre-emphasis
STFT
STFT
STFT
Magnitude Squared
Magnitude Squared
Magnitude Squared
Triangular Frequency Integration
Critical-Band Integration
Gammatone Frequency Integration
Equal Loudness Pre-emphasis
Medium-Time Power Bias Subtraction
Logarithmic Nonlinearity log(•)
Power Function Nonlinearity (•)^0.33
Power Function Nonlinearity (•)^0.1
DCT
IDFT
DCT
Levinson-Durbin Recursion
Cepstral Recursion
Mean Normalization
Mean Normalization
Mean Normalization
Output Features MFCC
Output Features PLP
Output Features PNCC

Figure 1: *Comparison of the PNCC feature extraction discussed in this paper with MFCC and PLP feature extraction.*

## 2. Derivation of the power function nonlinearity

Currently the most widely used feature extraction algorithms are Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). Both the MFCC and PLP procedures include intrinsic nonlinearities: PLP passes the amplitude-normalized short-time power of critical-band filters through a cube-root nonlinearity to approximate the power law of hearing [1, 4] while the MFCC procedure passes its filter outputs through a logarithmic function. Even though the importance of auditory nonlinearity has been confirmed in several studies (*e.g.* [5, 6]), there has been relatively little analysis concerning the effects of peripheral nonlinearities. In sophisticated auditory models such as [7], the curve relating input level in

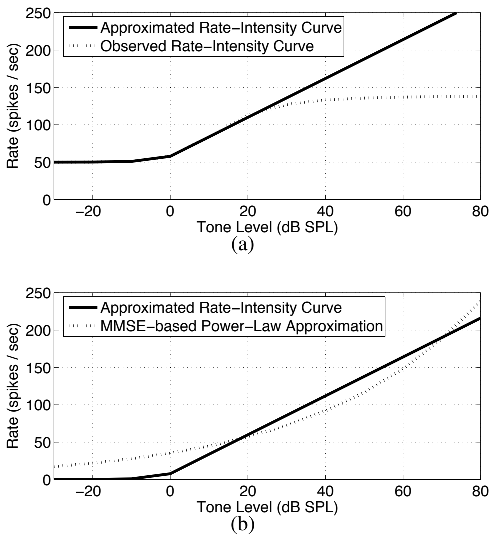


Figure 2: *Upper panel: Observed frequency-averaged mean rate of auditory-nerve firings versus intensity (dotted curve) and its piece-wise linear approximation (solid curve). Lower panel: Piece-wise linear rate-level curve with no saturation (solid curve) and best-fit power function approximation (dotted curve).*

decibels to the auditory-nerve firing rate is usually S-shaped. For example, the dotted line in the upper panel of Fig. 2 shows the relation between the intensity of a tone in dB and the rate of the auditory-nerve response, averaged across frequency, based on predictions by the model of [7] with the spontaneous rate of firing assumed to be 50 spikes/second. This curve is an abstract of results from many studies that observe that the firing rate is almost constant if the input SPL is smaller than a threshold intensity (-10 dB in this case), that the rate increases approximately linearly between 0 and 20 dB, and that it saturates at higher input levels. Because the logarithmic nonlinearity used in MFCC features does not exhibit threshold behavior, for speech segments of low power the output of the logarithm nonlinearity can produce large output changes even if the changes in input are small. This characteristic, which can degrade speech recognition accuracy, becomes very obvious as the input approaches zero, since even small differences in additive noise can produce large differences in the output of the nonlinearity. With a power-function nonlinearity, the output is close to zero if the input is very small, which is what is observed in human auditory processing.

The solid curve in the upper panel of Fig. 2 is a piecewise-linear approximation to the dotted curve in the same panel for intensities below 0 dB. For greater input intensities this solid curve is a linear approximation to the dynamic behavior of the rate-intensity curve between 0 and 20 dB. Hence, this solid curve exhibits threshold behavior but no saturation. We prefer to model the higher intensities with a curve that continues to increase linearly to avoid spectral distortion caused by the saturation seen in the dotted curve in the upper panel of Fig. 2.

The solid curve of the lower panel of Fig. 2 reprises the solid curve in the upper panel of the same figure, but translated downward so that for small intensities the output is zero (rather than the physiologically-appropriate spontaneous rate of 50 spikes/s). The dotted power function in that panel is the MMSE-based best-fit power function to the piecewise-linear solid curve. The reason for choosing the power-law nonlinearity instead of the solid curve in Fig. 2 is that the dynamic behavior of the output does not depend critically on the input amplitude. This nonlinearity, which is what is used in PNCC feature extraction, is described by the equation

$$y \;=\; x^{a_0} \tag{1}$$

with the best-fit value of the exponent observed to be $a_0 = 0.1$. We note that this exponent differs somewhat from the power-law exponent of $0.33$ used for PLP features, which is based on Steven's power law of hearing [4]. While our power-function nonlinearity may appear to be only a crude approximation to the physiological rate-intensity function, we will show in Sec. 4 that it provides substantial improvement in recognition accuracy compared to the traditional log nonlinearity used in MFCC processing. An attractive feature of the power-law nonlinearity is that the dynamic behavior of the output does not depend critically on the input amplitude, as in the case of MFCC's log nonlinearity.

## 3. Medium-duration power bias removal

In this section, we discuss medium-duration power normalization, which provides further decreases in WER. This operation is motivated by the fact that perceptual systems focus on target signal changes and largely ignore constant background levels. The algorithm presented in this section resembles conventional spectral subtraction in some ways, but instead of estimating noise power from non-speech segments of an utterance, we simply subtract a bias that is assumed to represent an unknown level of background stimulation.

### 3.1. Medium-duration power bias removal based on arithmetic-to-geometric mean ratios

Most speech recognition and speech coding systems use analysis frames of duration between 20 ms and 30 ms. Nevertheless, it is frequently observed that longer analysis windows provide better performance for noise modeling and/or environmental normalization, presumably because noise power changes more slowly than speech power. In PNCC processing we estimate the medium-duration power of speech signal $Q(i, j)$ by computing the running average of $P(i, j)$ , the power observed in a single analysis frame, according to the equation:

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

where $i$ represents the channel index and $j$ is the frame index. As mentioned before, we use a 25.6-ms Hamming window, and 10 ms between successive frames.

We found that $M = 3$ is optimal for speech recognition performance, which corresponds to seven consecutive windows or 85.6 ms. We find it convenient to use the ratio of arithmetic mean to geometric mean (the "AM-to-GM ratio") to estimate the degree of speech corruption. Because addition is easier to handle than multiplication and exponentiation to the power of $1/J$, we use the logarithm of the ratio of arithmetic and geometric means in the $i$-th channel as the normalization statistic:

$$G(i) = \log\left[\sum_{j=0}^{J-1} \max(Q(i, j), \epsilon)\right] - \frac{1}{J}\sum_{j=0}^{J-1} \log\left[\max\left(Q(i, j), \epsilon\right)\right] \tag{3}$$

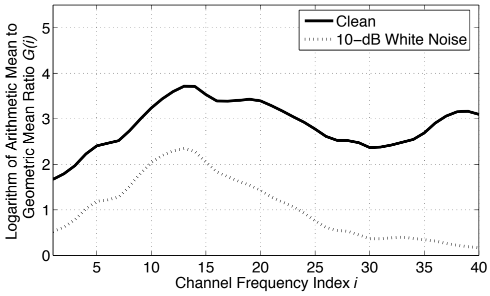


Figure 3: *Comparison between $G(i)$ coefficients for clean speech and speech in 10-dB white noise, using $M = 3$ in (2).*

The $\epsilon$ term in the above equation is imposed to avoid evaluations of negative infinity. Fig. 3 illustrates typical values of the statistic $G(i)$ for clean speech and speech that is corrupted by additive white noise at an SNR of 10 dB. As can be seen, values of $G(i)$ tend to decrease with increased noise level. $G(i)$ was estimated from the 1,600 utterances of the DARPA resource management training set, with $M = 3$ as in (2).

### 3.2. Removing the power bias

In this subsection, we explain how to estimate $B(i)$ for each channel of each test utterance using information from the clean training database. Power bias removal consists of estimating $B(i)$, the unknown level of background excitation in each channel and then computing the system output that would be obtained after it is removed.

If we could assume a value for $B(i)$, the normalized power $\tilde{Q}(i, j|B(i))$ is given by following equation:

$$\tilde{Q}(i, j|B(i)) = \max(Q(i, j) - B(i), d_0 Q(i, j)) \qquad (4)$$

In the above equation $d_0$ is a small constant (currently $10^{-3}$ that prevents $\tilde{Q}(i, j)$ from becoming negative. Using this normalized power $\tilde{Q}(i, j|B(i))$ , we can define the parameter $\tilde{G}(i|B(i))$ from (3) and (4):

$$\tilde{G}(i|B(i)) = \log\left[\sum_{j=0}^{J-1} \max\left(\tilde{Q}(i, j|B(i)), c_f(i)\right)\right] - \frac{1}{J}\sum_{j=0}^{J-1} \log\left[\max\left(\tilde{Q}(i, j|B(i)), c_f(i)\right)\right] \qquad (5)$$

The floor coefficient $c_f(i)$ is defined by:

$$c_f(i) = d_1\left(\frac{1}{J}\sum_{j'=0}^{J-1} Q(i, j')\right) \qquad (6)$$

In our system, we use $d_1$ of $10^{-3}$, causing $d_1$ to represent $-30$ dB of the channel average power. In our experiments, we observed that $c_f(i)$ plays a significant role in making the power bias estimate reliable, so its use is highly recommended. We noted previously that the $G(i)$ statistic is smaller for corrupt speech than it is for clean speech. From this observation, we can define the estimated power bias $B^*(i)$ as the smallest power which makes the AM-to-GM ratio the same as that of clean speech. This can be represented by the equation

$$B^*(i) = \min\left\{B(i)\middle|\tilde{G}(i|B(i)) \geq G_{cl}(i)\right\} \qquad (7)$$

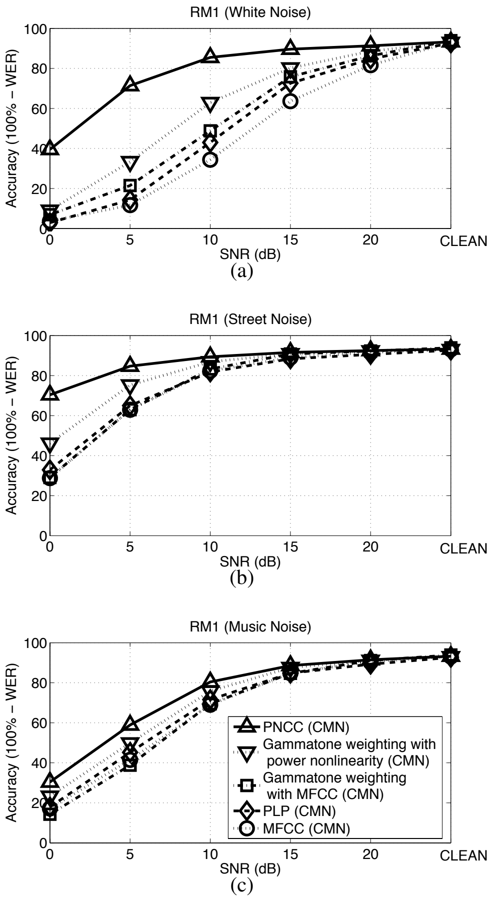


Figure 4: *Speech recognition accuracy obtained in different environments: (a) additive white gaussian noise, (b) noise recorded on an urban street, and (c) background music from the DARPA HUB4 database.*

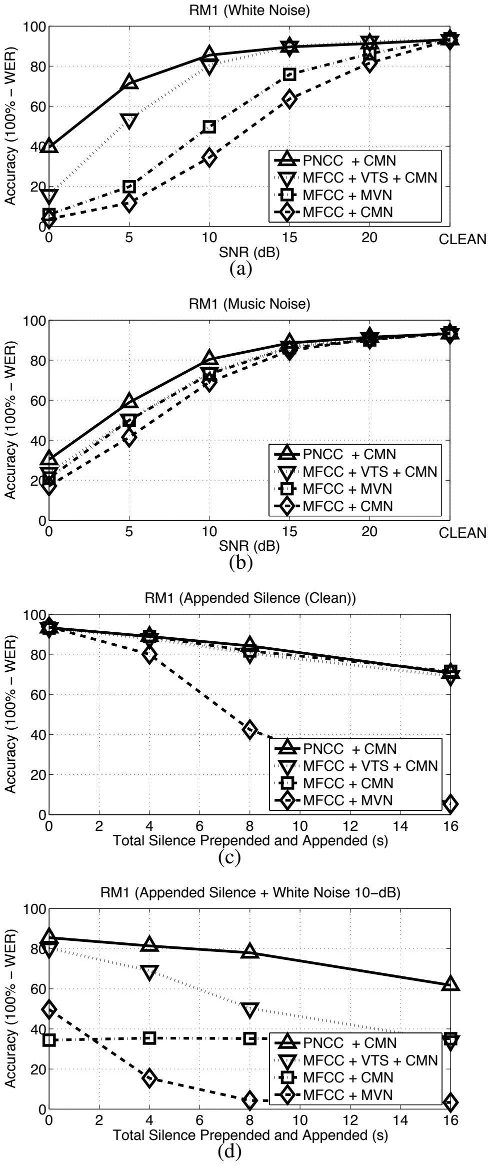
where $G_{cl}(i)$ is the value of $G(i)$ observed for clean speech, as shown in Fig. 3 Hence we obtain $B^*(i)$ by increasing $B(i)$ in steps from $-50$ dB relative to the average power in Channel $i$ until $\tilde{G}(i|B(i))$ becomes greater than $G_{cl}(i)$ as in Eq. (7). Using this procedure for each channel, we can obtain $\tilde{Q}(i, j|B^*(i))$. Thus, for each time-frequency bin represented by (i, j), the power normalization gain is given by:

$$w(i, j) = \frac{\tilde{Q}(i, j|B^*(i))}{Q(i, j)} \qquad (8)$$

For smoothing purposes, we average across channels from the $i-N$th channel up to the $i+N$th channel. Thus, the final power $\tilde{P}(i, j)$ is given by the following equation,

$$\tilde{P}(i, j) = \left(\frac{1}{2N+1}\sum_{i'=max(i-N,1)}^{min(i+N,C)} w(i', j)\right) P(i, j) \qquad (9)$$

where $C$ is total number of channels. In our algorithm, we use $N = 5$ and a total number of 40 gammatone channels. This normalized power $\tilde{P}(i, j)$ is applied to the power function nonlinearity as shown in the block diagram of Fig. 1.



(a)

(b)

(c)

(d)

Figure 5: *Speech recognition accuracy obtained in different environments: (a) additive white gaussian noise, (b) background music, (c) silence prepended and appended to the boundaries of clean speech, and (d) 10-dB of white Gaussian noise added to the data used in panel (c).*

## 4. Experimental results and conclusions

The PNCC system described in Secs. 2 and 3 was evaluated by comparing the recognition accuracy obtained using the CMU Sphinx 3.8 system with Sphinxbase 0.4.1 using PNCC with that of conventional MFCC processing, and with PLP processing as included in HTK 3.4. For training and testing, we used subsets of 1600 utterances and 600 utterances respectively from the DARPA Resource Management (RM1) database and trained using SphinxTrain 1.0. To evaluate the robustness of the feature extraction approaches we digitally added three different types of noise: white noise, street noise, and background music. The background music was obtained from a musical segment of the DARPA Hub 4 Broadcast News database, while the street noise was recorded by us on a busy street. We prefer to characterize improvement in recognition accuracy by the amount of lateral threshold shift provided by the processing. For white noise, PNCC provides an improvement of about 12 dB to 13 dB compared to MFCC, as shown in Fig. 4. For the street noise and the music noise, PNCC provides 8 dB and 3.5 dB shifts, respectively. These improvements are greater than improvements obtained with other current state-of-the-art algorithms such as Vector Taylor Series (VTS) [8], as shown in Fig. 5 We observe that if silence is added to the beginning and ends of the utterances, performance using some algorithms like mean-variance normalization (MVN) suffers if a good voice activity detector (VAD) is not included, as shown in Fig. 5. PNCC, on the other hand, degrades only slightly under the same conditions and without VADs.

Fig. 4 also demonstrates the amount of improvement provided by (1) the switch from the triangular MFCC filters to Gammatone filters, (2) the switch from the logarithmic nonlinearity to the power law nonlinearity, and (3) the use of medium-duration power bias removal. PNCC requires only slightly more computation than MFCC and much less computation than VTS. We also note that the use of the power nonlinearity and gammatone weighting with the DCT (dels in Fig. 4) still performs significantly better than PLP.

Open Source MATLAB code for PNCC can be found at http://www.cs.cmu.edu/˜robust/archive/algorithms/PNCC_IS2009.

The code in this directory was used for obtaining the results in this paper.

## 5. Acknowledgements

This research was supported by NSF (Grant IIS-0420866). The authors are thankful to Prof. Suryakanth Gangashetty and Hyongjoon Lim for helpful discussions.

## 6. References

[1] H. Hermansky, “Perceptual linear prediction analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87(4), no. 4, pp. 1738–1752, Apr. 1990.

[2] C. Kim and R. M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *INTERSPEECH-2008*, Sept. 2008.

[3] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, “Complex sounds and auditory images,” in *Auditory and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds. Pergamon Press, Oxford, 1992, pp. 429–446.

[4] S. S. Stevens, “On the psychophysical law,” *Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.

[5] Y.-H. Chiu and R. M. Stern, “Analysis of physiologically-motivated signal processing for robust speech recognition,” in *INTERSPEECH-08*, Sept. 2008, pp. 1000–1003.

[6] C. Kim, Y.-H. Chiu, and R. M. Stern, “Physiologically-motivated synchrony-based processing for robust automatic speech recognition,” in *INTERSPEECH-2006*, Sept. 2006, pp. 1975–1978.

[7] X. Zhang, M. G. Heinz, I. C. Bruce, and, L. H. Carney, “A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression,” *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 648–670, Feb 2001.

[8] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996, pp. 733–736.