다층 퍼셉트론의 층별 학습을 위한 중간층 오차 함수

A New Hidden Error Function for Layer-By-Layer Training of Multilayer Perceptrons

오상훈
목원대학교

Oh Sang-Hoon
Mokwon University

요약

다층 퍼셉트론의 학습을 빠르게 하기 위한 방법으로 층별 학습이 제안되었었다. 이 방법에서는 각 층별로 주어진 오차함수를 최적화 방법을 사용하여 감소시키도록 학습이 이루어진다. 이 경우 중간층 오차함수가 학습의 성능에 큰 영향을 미치는 데, 이 논문에서는 층별 학습의 성능을 개선하기 위한 중간층 오차함수를 제안한다. 이 중간층 오차함수는 출력층 오차함수에서 중간층 가중치의 학습에 관계된 성분을 유도하는 형태로 제안된다. 제안한 방법은 필기체 숫자 인식과 고립단어인식 문제의 시뮬레이션으로 효용성을 확인하였다.

Abstract

LBL(Layer-By-Layer) algorithms have been proposed to accelerate the training speed of MLPs(Multilayer Perceptrons). In this LBL algorithms, each layer needs a error function for optimization. Especially, error function for hidden layer has a great effect to achieve good performance. In this sense, this paper proposes a new hidden layer error function for improving the performance of LBL algorithm for MLPs. The hidden layer error function is derived from the mean squared error of output layer. Effectiveness of the proposed error function was demonstrated for a handwritten digit recognition and an isolated-word recognition tasks and very fast learning convergence was obtained.

I. 서 론

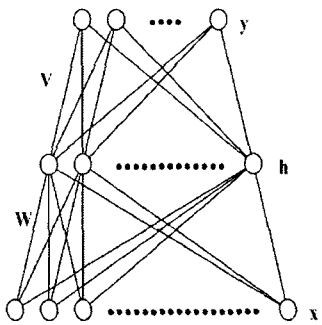
다층퍼셉트론(MLP: Multilayer Perceptron)은 충분한 수의 중간층 노드가 있으면 임의의 함수를 허용오차 내에서 근사화 할 수 있다는 특성 때문에 패턴인식, 시계열 예측, 비선형 제어, 통신 등에 응용되고 있다. MLP의 학습으로는 EBP(Error Back Propagation)가 가장 널리 사용되는 데, 이는 고정된 학습률을 사용한 강하 학습법(gradient descent)이다[1]. 이 방법은 학습속도가 느리기 때문에, 이에 대한 해결책으로 LBL(layer-by-layer) 최적화 방법이 제안되었는데, 이는 MLP의 각 계층을 비선형 부분과 선형 부분으로 나누어 학습을 진행시킨다[2,3]. 각 계층의 선형 부분은 최소 자승(least squares) 문제 형태로 풀어진다. 비록 이 방법이 컨쥬게이트 그레디언트(conjugate gradient)나 뉴튼(Newton) 방법 보다 훨씬 작은 계산량과 빠른 학습속도를 보이지만, 중간층에 설정되는 목표 값 때문에 학습이 안되는 문제에 직면하게 된다.

Yam과 Chow는 중간층의 목표 값 설정 때문에 학습이 정체되는 문제를 해결하기 위하여 LBL과 EBP의 접목 방법을 제안하였다[4]. Chen과 Manry도 출

력층은 최소 자승(least squares) 문제 형태의 학습을 적용하며 중간층은 EBP 형태의 학습을 제안하였다[5]. Ergezinger와 Thomsen은 중간층 노드의 시그모이드 함수를 선형적으로 근사화시킨 후 중간층 가중치를 학습시키는 형태의 LBL 방법을 제안하였다[6]. 이 방법들은 중간층 목표 값을 설정하지 않는 방법이지만, 휴리스틱(Heuristic)한 방법을 사용한다.

이 논문은 LBL 학습에서 출력층의 오차함수를 근거로 중간층의 가중치 학습에 관계된 성분을 유도한다. 이렇게 유도된 성분에 기반한 LBL 학습은 일반적인 LBL 학습의 정체 문제를 휴리스틱(Heuristic)한 밥법이나 최적화되지 않은 학습률을 사용함 없이 해결한다.

II. 다층퍼셉트론의 LBL 학습



▶▶ 그림 1. 다층퍼셉트론 구조

다층퍼셉트론(MLP)이 N개의 입력 노드와 H개의 중간층 노드 및 M개의 출력 노드들로 구성되어 있다고 하자. 어떤 입력패턴 $\boldsymbol{x}=[x_1, x_2, \ldots, x_N]$이 MLP에 입력되면, j번째 중간층 노드의 값은

$h_j = f(\widehat{h}_j) = \tanh(\widehat{h}_j/2),\ (j=2,1,\ldots,H)$

와 같이 주어진다. 여기서 $f(.)$ 는 중간층 노드의 비선형 함수이며 $\widehat{h}_j = \sum_{i=0}^{N} w_{ji}x_i$ 는 중간층 노드에 입력되는 가중치 합이다. w_{ji}는 x_i와 h_j를 연결하는 중간층 가중치이며 $x_0=1$로 주어지기에 w_{j0}는 바이어스라고 불린다. 같은 형태로 k번째 출력 노드 y_k에 입력되는 가중치 합은

$\widehat{y}_k = \sum_{j=0}^{H} v_{kj}h_j, (k=1,2,\ldots,M)$ 이고, v_{kj}는 h_j와 y_k를 연결하는 출력층 가중치이고, $h_0=1$이며 v_{k0}는 바이어스이다.

P개의 학습패턴 $\boldsymbol{x}^{(p)}(p=1,2,\ldots,P)$와 이들의 출력층 목표벡터 $\boldsymbol{t}^{(p)}$가 주어지면, 가중치들은 출력층에서

$$E^{out} = \frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{M}(t_k^{(p)} - y_k^{(p)})^2 \qquad (1)$$

로 주어지는 MSE(mean squared error)를 최소화시키도록 변경된다. 이 논문은 출력층 노드가 선형이라고 가정하여 $y_k^{(p)} = \widehat{y}_k^{(p)}$이고 $t_k^{(p)} = \widehat{t}_k^{(p)}$로 둔다.

다층퍼셉트론의 LBL 학습에서 한 학습 epoch은 3개의 독립적인 최소 자승(least squares) 문제로 분리된다[6]. 최소 자승(Least squares) 문제는 2차 함수 형태를 지니므로, 강하 학습(gradient descent) 알고리즘이 최적의 학습률을 지니도록 할 수 있다. 다음의 3 단계가 최적 학습률을 지닌 LBL 학습을 요약하여 설명한다.

- 단계 1: $W=\{w_{ji}\}$가 고정되어 있고 $\boldsymbol{t}^{(p)}$가 주어진 상황에서, E^{out}이 최소값을 지니도록

$$\Delta v_{kj} = -\eta_k^{out}\frac{\partial E^{out}}{\partial v_{kj}} = \eta_k^{out}\sum_{p=1}^{P}(t_k^{(p)} - y_k^{(p)})h_j^{(p)} \qquad (2)$$

와 같이 학습한다. 여기서, η_k^{out} 은 y_k 에 연관된 출력층 가중치들의 학습률로써 그 최적값은

$$\eta_k^{out} = \frac{\sum_{j=0}^{H}\left(\frac{\partial E^{out}}{\partial v_{kj}}\right)^2}{\sum_{p=1}^{P}\left(\sum_{j=0}^{H}\frac{\partial E^{out}}{\partial v_{kj}} h_j^{(p)}\right)^2}, k=1,2,\ldots,M \quad (3)$$

로 구해진다.

• 단계 2: 단계 1에 의해 변경된 V를 사용하여 중간층 노드의 목표값을

$$z_j^{(p)} = h_j^{(p)} + \zeta_p \beta_j^{(p)} \quad (4)$$

와 같이 정한다. 여기서,

$$\beta_j^{(p)} \equiv -\frac{\partial E^{out}}{\partial h_j^{(p)}} = \sum_{k=1}^{M}\left(t_k^{(p)} - y_k^{(p)}\right) v_{kj} \quad (5)$$

이고, ζ_p는 가상의 중간층 목표값을 할당하기 위한 학습률이다. v_{kj}가 고정되었다고 가정하면, ζ_p의 최적값이

$$\zeta_p = \frac{\sum_{j=1}^{H}\left(\beta_j^{(p)}\right)^2}{\sum_{k=1}^{M}\left(\sum_{j=1}^{H} v_{kj}\beta_j^{(p)}\right)^2}, p=1,2,\ldots,P \quad (6)$$

와 같이 구해진다. 식 (6)을 이용하여 식 (4)와 같이 가상의 중간층 목표값이 정해진 후, $-1 < z_j^{(p)} < 1$을 만족하도록 절삭의 과정을 거친다. 그 다음 $h_j^{(p)}$에 대해 입력되는 가중치 합의 목표값이

$$\hat{z}_j^{(p)} = f^{-1}(z_j^{(p)}) = 2\tanh^{-1}(z_j^{(p)}) \quad (7)$$

와 같이 구해진다.

• 단계 3: 단계 2에서 구한 $\hat{z}_j^{(p)}$와 주어진 학습 패턴 $\boldsymbol{x}^{(p)}$를 활용하여,

$$E^{hid} = \frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{H}\left(\hat{z}_j^{(p)} - \hat{h}_j^{(p)}\right)^2 \quad (8)$$

을 최소화 시키도록 $W=\{w_{ji}\}$를

$$\Delta w_{ji} = -\eta_j^{hid}\frac{\partial E^{hid}}{\partial w_{ji}} \quad (9)$$

와 같이 변경한다. 여기서, 학습률 η_j^{hid}의 최적값은

$$\eta_j^{hid} = \frac{\sum_{i=0}^{N}\left(\frac{\partial E_n^{hid}}{\partial w_{ji}}\right)^2}{\sum_{p=1}^{P}\left(\sum_{i=0}^{N}\frac{\partial E_n^{hid}}{\partial w_{ji}} x_i^{(p)}\right)^2} \quad (10)$$

로 주어진다.

단계 1, 2, 3은 모두 선형 문제를 최적 학습률을 사용하여 푸는 형태이므로 빠른 수렴성을 보인다. 그렇지만, 단계 2에서 주어지는 중간층 노드의 목표값들이 선형적으로 분리될 수 없으면, 단계 3에서 중간층 가중치의 학습에 의해 E^{hid}를 충분히 줄이는 것은 불가능하다. 이것이 MLP의 LBL 학습에서 더 이상 오차가 줄지 않는 학습의 정체 문제를 일으키게 된다.

III. LBL 학습을 위한 중간층 오차함수

식 (8)과 같이 주어진 중간층 노드의 가중치 합에 대한 오차함수는 중간층 가중치의 학습에 대한 근거를 제공한다. 그렇지만, 중간층 노드의 가중치 합은 시그모이드 함수를 거친 후 출력층 노드로 전달되며, 학습은 출력층 노드의 오차를 줄이는 것이 최종 목적이므로 중간층의 오차함수는 출력층의 오차함수로부터 유도되어져야 한다.

이를 위하여, 중간층 노드가 목표값 $z_j^{(p)}$를 지닌 경

우의 출력값을 $Y_k^{(p)}$라고 두면

$$Y_k^{(p)} = \sum_{j=0}^{H} v_{kj} z_j^{(p)} \tag{11}$$

이고, 출력층의 오차함수는

$$E^{out} = \frac{1}{2} \sum_{p=1}^{P} \sum_{k=1}^{M} [(t_k^{(p)} - Y_k^{(p)}) + (Y_k^{(p)} - y_k^{(p)})]^2 \tag{12}$$

와 같이 둘 수 있다. 이 수식을 전개한 후, $(t_k^{(p)} - Y_k^{(p)})$가 $\sum_{j=0}^{H} v_{kj}(z_j^{(p)} - h_j^{(p)})$와 상관관계가 없다고(uncorrelated) 가정하면

$$E^{out} \approx \frac{1}{2} \sum_{p=1}^{P} \sum_{k=1}^{M} (t_k^{(p)} - Y_k^{(p)})^2 + \frac{1}{2} \sum_{p=1}^{P} \sum_{k=1}^{M} [\sum_{j=0}^{H} v_{kj}(z_j^{(p)} - h_j^{(p)})]^2 \tag{13}$$

와 같이 근사화 된다. 식 (13)에서 두 번째 항이 중간층 노드의 값에 관련되어 있으므로 E_n^{hid}로 두면, 이 E_n^{hid}을 근거로 중간층 노드의 오차함수를 유도하겠다.

먼저 $(z_j^{(p)} - h_j^{(p)})$ 가 1계 미분 가능한 함수이므로, 1차 Taylor 급수 근사화를 이용하면

$$(z_j^{(p)} - h_j^{(p)}) \approx f'(\hat{h}_j^{(p)})(\hat{z}_j^{(p)} - \hat{h}_j^{(p)}) \tag{14}$$

와 같이 된다. 이를 식 (13)의 두 번째 항에 대입한 후 중간층 노드 간에 $(z_j^{(p)} - h_j^{(p)})$들이 상관관계가 없다는(uncorrelated) 가정을 적용하면

$$E_n^{hid} \approx \frac{1}{2} \sum_{p=1}^{P} \sum_{j=0}^{H} \sum_{k=1}^{M} v_{kj}^2 [f'(\hat{h}_j^{(p)})]^2 (\hat{z}_j^{(p)} - \hat{h}_j^{(p)})^2 \tag{15}$$

와 같이 유도된다. 여기서, $\sum_{k=1}^{M} v_{kj}^2$ 이 거의 상수에 가깝다고 가정하고 또한 $f'(\hat{h}_j^{(p)}) \approx f'(\hat{z}_j^{(p)})$ 라고 두면

$$E_n^{hid} = \frac{1}{2} \sum_{p=1}^{P} \sum_{j=1}^{H} (\hat{z}_j^{(p)} - \hat{h}_j^{(p)})^2 [f'(\hat{z}_j^{(p)})]^2 \tag{16}$$

와 같이 된다. 이 오차함수에서 마지막 항은 중간층 노드가 시그모이드 함수의 선형영역에 있는 지 혹은 포화 영역에 있는 지에 따라 중간층 가중치의 변경량을 조절하는 역할을 한다. 이는 MLP의 출력층에서 MSE를 최소화 시키는 것이 EBP 학습의 목적이므로, $\hat{z}_j^{(p)}$와 $\hat{h}_j^{(p)}$의 거리 보다는 $z_j^{(p)}$와 $h_j^{(p)}$의 거리를 고려하여 중간층 가중치가 변경되도록 하는 것이다.

E_n^{hid}를 최소화시키기 위하여 중간층 가중치는

$$\Delta w_{ji} = - \eta_j^{hid} \frac{\partial E_n^{hid}}{\partial w_{ji}} \tag{17}$$

에 따라 변경될 것이다. 여기서, η_j^{hid}는 학습률이고

$$\frac{\partial E_n^{hid}}{\partial w_{ji}} = - \sum_{p=1}^{P} (\hat{z}_j^{(p)} - \hat{h}_j^{(p)}) [f'(\hat{z}_j^{(p)})]^2 x_i^{(p)} \tag{18}$$

이다. 식 (17)에 따라 변경된 중간층 가중치를 식 (16)에 대입하면

$$E_n^{hid}(\eta_j^{hid}) = \frac{1}{2} \sum_{p=1}^{P} \sum_{j=1}^{H} (\hat{z}_j^{(p)} - \hat{h}_j^{(p)} + \eta_j^{hid} \sum_{i=0}^{N} \frac{\partial E_n^{hid}}{\partial w_{ji}} x_i^{(p)})^2 [f'(\hat{z}_j^{(p)})]^2 \tag{19}$$

이 된다. 따라서, 최적의 $\eta_j^{hid} (j = 1, 2, ..., H)$는 조건

$\partial E_n^{hid}(\eta_j^{hid}) / \partial \eta_j^{hid} = 0$ 에 의해

$$\eta_j^{hid} = \frac{\sum_{i=0}^{N}\left(\frac{\partial E_n^{hid}}{\partial w_{ji}}\right)^2}{\sum_{p=1}^{P}\left(\sum_{i=0}^{N}\frac{\partial E_n^{hid}}{\partial w_{ji}} x_i^{(p)}\right)^2 \left[f'(\hat{z}_j^{(p)})\right]^2} \quad (20)$$

와 같이 구해진다. 식 (17)-(20)이 단계 2의 수식을 대체한 것이며 이 논문에서 새롭게 제안된 학습 방법이다.

식 (4)에서 $\zeta_p \beta_j^{(p)}$ 가 작은 값을 지닐 경우, 테일러 급수 전개에 의해 식 (17)은

$$\Delta w_{ji} \approx \eta_j^{hid} \sum_{p=1}^{P} \zeta_p \delta_j^{hid}(\boldsymbol{x}^{(p)}) x_i^{(p)} \quad (21)$$

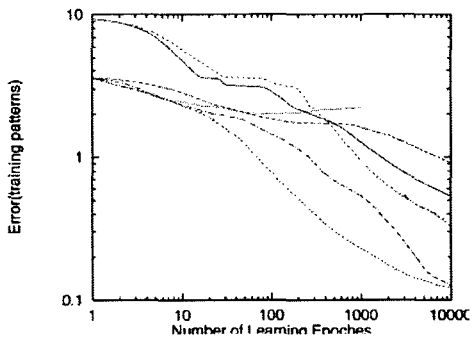
와 같이 근사화 된다.

IV. 시뮬레이션

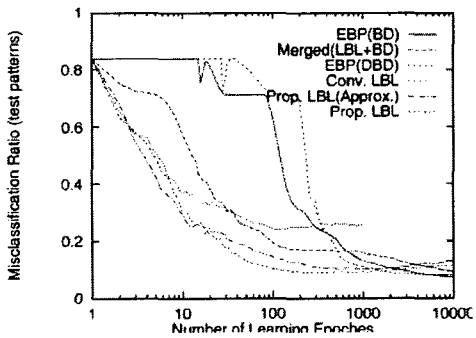
제안한 방법의 효용성을 검증하기 위하여 필기체 숫자 인식 문제를 시뮬레이션하였다. CEDAR 데이터베이스[7]에서 5000개의 숫자 영상을 추출하여 크기 정규화 과정을 거친 후 학습에 사용하였다. 숫자 영상의 크기는 12×12 픽셀이며 각 픽셀은 16가지 레벨의 값을 지닌다. MLP는 입력 144, 중간층 30, 출력층 10개의 노드들로 구성되었다.

MLP의 학습방법으로는 BD(Bold Driver)[5]를 이용한 EBP, DBD(Delta-Bar-Delta)를 이용한 EBP, 출력층은 LBL에 따른 학습을 수행하며 중간층은 EBP의 BD에 따른 접목방법, 일반적인 LBL 방법, 그리고, 제안한 LBL 방법과 수식 (21)로 주어진 근사화 방법을 사용하였다.

[그림 2(a)]는 학습패턴에 대한 출력층의 MSE를 보여준다. BD와 DBD 방법은 학습률이 최적 값이 아니므로 학습속도가 느리다. 일반적인 LBL 방법은 비록 학습 초기에 MSE가 빠르게 감소하였지만 곧 학습의 정체현상이 나타났다. 비록 접목 방법에서는 학습의 정체 현상이 나타나지 않지만 학습 속도가 느리다. 이에 반하여, 제안한 LBL 방법은 빠른 학습속도를 보여준다. 그리고, 제안한 방법을 근사화시킨 경우 시뮬레이션 결과는 학습 속도가 다소 느려졌지만 다른 방법들 보다는 빠름을 볼 수 있다. [그림 2(b)]는 2213개의 학습시키지 않은 시험패턴에 대한 오인식률을 보여준다. 학습패턴에 대한 MSE의 경우와 마찬가지로 EBP에 근거한 방법들은 오인식률이 천천히 감소함을 볼 수 있다. 그렇지만, 제안한 방법은 오인식률이 빠르게 감소하였다. 비록 학습이 진행 될수록 시험패턴에 대한 오인식률이 증가하지만, 이것은 학습패턴에 대한 학습이 필요 이상으로 많이 되어 시험 패턴에 대한 특성은 오히려 나빠지기 때문이다.



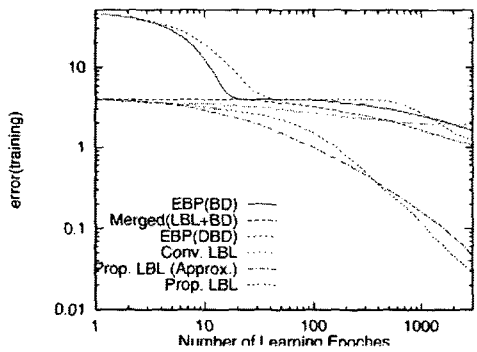
(a) 학습패턴에 대한 MSE



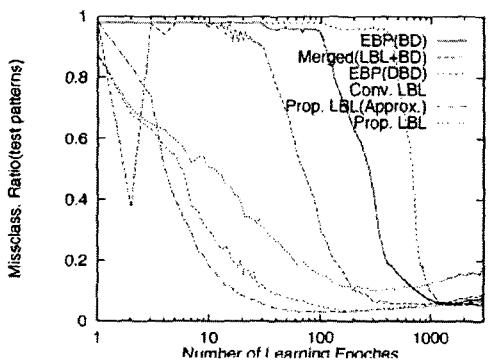
(b) 시험패턴에 대한 오인식률

▶▶ 그림 2. 필기체 숫자 인식 시뮬레이션

더욱 더 철저한 효용성 검증을 위하여 고립단어 인식 문제도 시뮬레이션하였다. 50단어가 인식대상이며, 900개의 학습패턴에서 1024차원의 ZCPA 특징[8]을 추출 후 50개의 중간층 노드를 지닌 MLP에 학습시켰다. [그림 3]은 학습패턴에 대한 MSE와 1050개의 시험패턴에 대한 오인식률을 나타내는데, [그림 2]에서와 마찬가지로 제안한 방법이 다른 방법보다 훨씬 더 학습속도가 빠름을 알 수 있다. 이 그림에서도 [그림 2(b)]에서와 마찬가지로 학습이 진행될수록 시험패턴에 대한 오인식률이 증가하는 현상이 나타난다. 이는 학습 패턴에 대한 과도한 학습이 시험패턴에 대한 특성을 저하시키기 때문이다.



(a) 학습패턴에 대한 MSE



(b) 시험패턴에 대한 오인식률

▶▶ 그림 3. 고립단어 인식문제 시뮬레이션

V. 결 론

이 논문에서는 MLP의 학습속도를 향상시키는 방안으로 사용되는 LBL에서 학습의 정체 현상을 해결하기 위하여 중간층 오차함수를 제안하였다. 제안한 중간층 오차함수는 출력층의 MSE 함수로부터 중간층 가중치의 학습에 관련된 부분 만을 유도하는 형태로 주어졌다. 이 새로운 중간층 오차함수는 중간층 가중치의 변경량이 중간층 노드 값의 시그모이드 함수에 위치한 영역에 따라 적절히 변하도록 하는 성질을 지녔다. 필기체 숫자 인식과 고립단어 인식 문제의 시뮬레이션으로 제안한 방법을 다른 방법들과 비교하였는데, 주장한 바와 같이 제안한 방법이 학습초기부터 MSE가 급격히 줄어듦을 확인할 수 있었다.

▮참 고 문 헌▮

[1] D. E. Rumelhart and J. L. McClelland, Parallel Distributed Processing. MIT Press, Cambridge, MA, 1986.

[2] R. Parisi, E. D. Di Claudio, G. Orlan야, and B. D. rao, "A generalized learning paradigm exploiting the structure of feedforward neural networks," IEEE Trans. Neural Networks, Vol.7, pp.1450-1459, 1996.

[3] G.-J. Wang and C.-C. Chen, "A fast multilayer neural networks training algorithm based on the layer-by-layer optimizing procedures," IEEE Trans. Neural Networks, Vol.7, pp.768-775, 1996.

[4] J. Y .F. Yam and W. S. Chow, "Extended least squares based algorithm for training feedforward networks," IEEE Trans. Neural Networks, Vol.8, pp.806-810, 1997.

[5] H.-H. Chen, M. T. Manry, and H. Chandrasekaran, "A neural network training algorithm utilizing multiple set of linear equations," Neurocomputing, Vol.25, pp.55-72, 1999.

[6] S. Ergezinger and E. Thomsen, "An accelerated learning algorithm for multilayer perceptrons: Optimization layer by layer," IEEE Trans. Neural Networks, Vol.6, pp.31-42, 1995.

[7] J. J. Hull, "A database for handwritten text recognition research," IEEE Trans. Pat. Ana. Mach. Int., Vol.16, No.5, pp.550-554, May 1994.

[8] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Trans. Speech and Audio Processing, Vol.7, pp.55-69, 1999.