한국음향학회지 제21권 제6호 pp. 522~530 (2002)

신경 회로망을 이용한 음성 신호의 장구간 예측

Long-term Prediction of Speech Signal Using a Neural Network

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보통신대학 전자공학부
(접수일자: 2002년 2월 5일; 채택일자: 2002년 7월 31일)

본 논문에서는 선형 예측 후에 얻어지는 잔차 신호 (residual signal)를 신경 회로망에 바탕을 둔 비선형 예측기로 예측하는 방법을 제안하였다. 신경 회로망을 이용한 예측 방법의 타당성을 입증하기 위해, 먼저 선형 장구간 예측기와 신경 회로망이 도입된 비선형 장구간 예측기의 성능을 서로 비교하였다. 그리고 비선형 예측 후의 잔차 신호를 양자화하는 과정에서 발생하는 양자화 오차의 영향에 대해 분석하였다. 제안된 신경망 예측기는 예측 오차뿐만 아니라 양자화의 영향을 함께 고려하였으며, 양자화 오차에 대한 강인성을 갖게 하기 위하여 쿤-터커 (Kuhn-Tucker) 부등식 조건을 만족하는 제한 조건 역전파 알고리즘을 새로이 제안하였다. 실험 결과, 제안된 신경망 예측기는 제한 조건을 갖는 학습 알고리즘을 사용했음에도 불구하고, 예측 이득이 크게 뒤떨어지지 않는 성능을 나타내었다.

핵심용어: 음성 신호 비선형 예측, 신경 회로망, 제한 조건을 갖는 역전파 알고리즘, 쿤-터커 부등식 조건
투고분야: 음성처리 분야 (2.4)

This paper introduces a neural network (NN) -based nonlinear predictor for the LP (Linear Prediction) residual. To evaluate the effectiveness of the NN-based nonlinear predictor for LP-residual, we first compared the average prediction gain of the linear long-term predictor with that of the NN-based nonlinear long-term predictor. Then, the effects on the quantization noise of the nonlinear prediction residuals were investigated for the NN-based nonlinear predictor. A new NN predictor takes into consideration not only prediction error but also quantization effects. To increase robustness against the quantization noise of the nonlinear prediction residual, a constrained back propagation learning algorithm, which satisfies a Kuhn-Tucker inequality condition is proposed. Experimental results indicate that the prediction gain of the proposed NN predictor was not seriously decreased even when the constrained optimization algorithm was employed.

Keywords: *Nonlinear prediction of speech signal, Neural network, Constrained back propagation algorithm, Kuhn-Tucker inequality condition*

ASK subject classification: *Speech signal processing (2.4)*

책임저자: 이기승 (kseung@kkucc.konkuk.ac.kr)
143-701 서울특별시 광진구 화양동 1번지
건국대학교 정보통신대학 전자공학부 1417호
(전화: 02-450-3489; 팩스:02-3437-5235)

I. 서 론

선형 예측 부호화 (LPC; Linear Predictive Coding)[1]는 음성 신호의 압축 기법으로 널리 사용되고 있는 기술이다. 표준화 알고리즘으로 널리 사용되는 CELP (Coded Excited Linear Prediction) 기법[2] 또한 선형 예측 부호화에 바탕을 둔 압축 기법이다. CELP에는 크게 두가지 예측 기법이 적용되었는데, 하나는 성도 전달 함수 (vocal tract transfer function)의 특성을 모델링하는 단구간 예측기 (short-term predictor)이며, 다른 하나는 단구간 예측 후의 잔차 신호, 즉 여기 신호 (excitation signal)를 예측하기 위한 장구간 예측기 (long-term predictor)이다. 단구간 예측기는 음성 신호를 선형 시변 필터로 모델링하는 이론[1]에 바탕을 둔 것으로, 음성 신호가 갖는 단구간 상관 특성을 이용한 것이다. 여기 신호에 대한 예측은 여기 신호가 갖는 펄스 성분을 효과적으로 제거하기 위해 피치와 유사한 간격만큼 떨어진 거리의 샘플들로 예측을 하는 장구간 예측이 사용된다.

여기 신호에 대한 장구간 예측은 단구간 예측보다 낮은 예측 이득 (Prediction gain)을 갖는데, 이의 주된 이유로는 앞단의 단구간 예측에 의해 선형 예측으로 표현될 수 있는 성분이 다소간 제거되었다는 점이다[3,4]. 또 다른 이유로는 선형 예측 후의 잔차 신호가 갖는 펄스 성분이 동일한 간격으로 나타나지 않고, 비선형적인 특징을 갖는데 있다. 이러한 사실을 입증하는 연구로서 Thyssen 등은 복수 개의 선형 예측기를 다단으로 연결하여도 여기 신호의 펄스 성분이 여전히 존재함을 보고하였다[3]. 이러한 이유로 현재의 음성 코덱에서는 장구간 예측시 분석 구간을 짧게 하여 비교적 적은 펄스가 포함되도록 하고 있다[2].

선형 시변 필터 (linear time-varying filter)가 비선형 필터의 일종인 시불변 쌍선형 (bilinear) 필터에 의해 모델링될 수 있음을 나타내는 쉐젠 이론 (Schetzen Theorem)[7]은, 비선형 예측기에 의해 여기 신호를 효과적으로 예측될 수 있음을 의미하기도 한다. 즉, 여기 신호의 펄스 성분이 시변 특성을 갖는다라고 가정할 때, 여기 신호는 시불변 비선형 예측기에 의해 모델링될 수 있기 때문이다. 이에 따라 다양한 형태의 비선형 예측기를 사용하여 선형 예측 후의 잔차 신호, 즉 여기 신호를 예측하는 연구가 시도되었다[3-6,8,9].

볼테라 필터 (Volterra filter)와 신경망을 이용하여 여기 신호를 예측하는 방법은 선형 예측기만을 사용한 경우와 비교하여, 2~3 dB정도의 예측 이득을 얻을 수 있음이 보고되었으며[3], Wu 등은 실제 비선형 예측기를 사용하여 CELP와 동일한 구조를 갖는 음성 부호화기를 제안하였다[5,6]. 여기서 사용한 비선형 예측기는 Thyssen의 연구와 마찬가지로 신경 회로망이 사용되었는데, 기존의 연구와 다른 점은 출력값을 다시 신경망의 입력 변수로 사용하는 회귀 신경망 (recurrent neural network)이 사용되었다는 점이다. Wu가 제안한 비선형 CELP 기법에서는 단구간 예측과 장구간 예측에 모두 신경망에 바탕을 둔 비선형 예측기가 사용되었으며, 기존의 4800 bps의 전송률에서 CELP 부호화기와 비교하여 객관적인 면에서 1.5 dB의 예측 이득 향상을 가져오며, 주관적인 면에서 MOS test상 0.35 향상된 값을 얻은 것으로 보고되었다[6].

한편 Mumolo 등의 연구에서는 비선형 예측기가 선형 예측기와 비교하여 항시 성능 우위를 가져오지는 않는다고 주장하였으며, 비선형 예측기와 선형 예측기를 적응적으로 사용하는 기법을 제안하였다[8]. 이 기법에서는 선형 예측과 비선형 예측을 동시에 수행하고 잔차 신호의 크기가 작아지는 예측기를 택하도록 하였다. 비선형 예측기는 Thyssen이 제안한 것과 유사한 볼테라 필터가 사용되었다. Maria 등의 연구에서는 비선형 예측기법으로, 원형기준함수 (radial basis function)를 사용하였는데, 신경망을 사용한 기존의 방법과 거의 유사한 성능을 얻었는 것으로 보고하였다[9].

이와 같은 기존의 비선형 예측에 관한 연구를 종합하면 비선형 예측기가 선형 예측기와 비교하여 객관적, 주관적인 성능 향상을 가져옴이 입증되었으나, 구현시의 복잡성과 안정도면에서의 취약성이 문제점으로 지적되고 있다[4-6].

구현시의 복잡성은 볼테라 필터의 예측 계수를 구하는 경우, 선형 예측 계수 추정의 2차원 확장된 형태로 나타나며, 결국은 제곱으로 증가된 계산량을 가져오는 것으로 설명될 수 있다. 신경 회로망을 사용하는 비선형 예측기의 경우, 역전파 알고리즘 (back propagation)[10]의 수렴 정도에 따라 계산량이 변동되는데, 수렴 상수가 작은 경우에는 계산 시간이 증가되며 수렴 상수가 큰 경우에는 계산 시간이 다소 감소하나, 수렴된 계수가 전역 최소 (global minimum)를 보장하지 못한다는 문제가 있다. 이러한 신경망의 문제점은 Thyssen 등에 의해 제안된 벡터 양자화-신경망 기법 (VQNN; Vector Quantized Neural Network)[4]에 의해 어느 정도 해결된다. 이 기법은 학습 과정에서 몇 개의 신경망을 미리 구성하고 여기 신호가 입력되면, 구성된 신경망 중에 가장 낮은 예측 오차를 갖는 신경망을 선택하는 것이다. 음성 신호의 압축과 같은

응용에서는 예측 오차의 크기뿐 아니고 신경망 변수를 사용하는데 필요한 비트수 또한 중요한 요소이므로, 제한된 개수의 신경망 변수를 사용하는 벡터양자화-신경망 기법은 매우 유용한 기법이라 하겠다. 본 논문에서도 음성 부호화의 적용을 고려하여 벡터양자화를 기반으로 하는 신경망 예측기를 사용하였다.

비선형 예측기의 안정성은 일반적인 비선형 필터의 역필터 (inverse filter)가 제한된 크기를 갖는 입력 신호 (bounded input)에 대해 제한된 범위내의 출력 신호 (bounded output)를 보장할 수 있는가로 설명된다. 자기회귀 (AR; Auto-Regressive) 모델을 사용하는 선형 예측기의 경우에는, 역필터의 극점 (pole)에 따라 시스템의 안정성이 결정되나[1], 일반적인 비선형 필터의 경우에는 안정성 정도를 정량화하여 나타내기 매우 어렵다[6]. 이러한 이유로 Wu 등의 연구에서는 안정성의 정도를 민감도 (sensitivity parameter)라는 변수를 도입하여 간접적으로 표현하는 기법을 사용하였다[6]. 폐회로 (closed loop)의 구조를 갖는 예측 부호화기의 경우, 부호화단에서는 원 신호를 알고 있기 때문에 예측도 시간 지연된 원 신호들을 통해 이루어질 수 있지만, 복호화단에서는 양자화 오차를 포함하는 복원된 신호만 알고 있으므로 예측기도 복원 신호에 의해 이루어진다. 여기서 민감도란 복원된 신호와 원 신호간의 차이가 발생하는 경우, 예측된 값이 실제 값과 어느 정도의 차이를 가져오는지를 나타내는 변수이다. 따라서 민감도가 큰 경우에는 예측 이득이 높더라도 실제 양자화 오차로 인한 영향을 매우 크게 받게 되어 전체적인 부호화 성능은 떨어지게 된다.

Wu 등의 연구에서는 음성 신호의 합성시, 신경망에 입력되는 잔차 신호를 학습 과정을 통해 유한 개의 코드로 제한시킴으로서, 신경망의 민감도에 덜 영향을 받도록 하였다[6]. 이와 같은 방법은 신경망의 민감도에 관계없이 최적의 예측 이득을 얻기 위해서는 바람직한 구조이지만, 잔차 신호를 유한 개로 표현함으로서 보다 다양한 음성 신호의 표현이 어려우며 잔차 신호의 부호화 기법이 벡터 양자화 방법으로 제한된다는 단점이 있다.

본 논문에서는 신경망 회로의 학습시 사용되는 역전파 알고리즘에 제한 조건을 두어 민감성을 억제시키는 기법을 제안하였다. 주어진 신경망 회로로부터 민감성을 먼저 유도하고 민감성과 예측 오차를 동시에 최소화하는 것은 불가능하므로, 민감성이 주어진 값보다 작아지도록 신경망을 학습하였다. 이와 같은 학습 방법은 제한 조건을 갖는 역전파 알고리즘에 의해 구현되며, 제한 조건을 갖는 역전파 알고리즘은 쿤-터커 부등식 조건 (Kuhn-Tucker inequality condition)[11]이 포함된 반복 추정식으로 나타내었다. 실험에서는 이와 같은 제한 조건이 포함된 경우의 신경망 회로의 예측 성능이 어떻게 변화하는지를 살펴보고, 민감도에 따른 예측 이득의 변화를 분석하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 선형 예측 후 잔차 신호를 신경망 회로에 바탕을 둔 비선형 예측기로 예측하는 방법에 대해 소개하며, 3장에서는 벡터 양자화 기법과 결합된 신경망 예측 기법을 알아본다. 4장에서는 신경망 예측기의 민감도를 분석하는 방법을 제시하며, 5장에서는 분석된 민감도를 억제시키는 제안된 신경망 예측기의 역전파 알고리즘을 제안한다. 6장에서는 실제 음성 신호에 적용된 예를 통해 제안된 예측기의 성능을 평가하며, 마지막 7장의 결론을 통해 본 논문을 맺는다.

II. 신경망을 이용한 잔차 신호의 예측

선형 예측 후에 얻어지는 잔차 신호는 준주기적인 펄스 성분을 포함하며, 이러한 펄스 성분은 장구간 예측기에 의해 제거될 수 있다[2]. 그러나 분석 구간의 크기가 커짐에 따라 펄스 성분은 비선형적인 특성이 나타나게 되어 장구간 선형 예측에 의해서도 여전히 펄스 성분이 남게 된다.

이러한 비선형적인 특성을 모델링하기 위해 본 논문에서는 비선형 예측기로 장구간 예측기를 설계하였다. 비선형 예측기로는 신경망[3-6], 2차 볼테라 필터[3,4,8], 원형 기준 함수 (RBF; radial basis function)[9] 등이 있으며, 이중 신경망을 이용한 예측기는 비선형 특성을 정형화된 식에 의해 표현하지 않으므로 여러 형태의 비선형 특성을 나타낼 수 있으며, 합성 필터 (synthesis filter)와 분석 필터 (analysis filter)의 안정성이 보장된다는 장점이 있다[3-6].

본 논문에서 구성한 비선형 장구간 예측기의 구조가 그림 1에 나타나 있다. 그림에서 $e(n)$은 예측하고자 하는 신호, 즉 선형 예측 후에 얻어지는 여기 신호를 나타내며 $\hat{e}(n)$은 예측된 여기 신호를 나타낸다. 그림 1의 신경망 예측기는 시간 지연된 샘플들로 입력 신호를 구성하는 시간 지연 신경망 (time-delay neural network)의 구조를 갖는다. 그림에서 f는 신경망에 포함된 비선형 함수로서 시그모이드 (sigmoid) 함수, 또는 쌍곡선 탄젠트 함수 (tanh)가 사용되는데 본 논문에서는 $\tanh(x) = \frac{1-\exp[-\alpha x]}{1+\exp[-\alpha x]}$를 사용하였다. 예측된 여기 신호 $\hat{e}(n)$은 다음과 같이

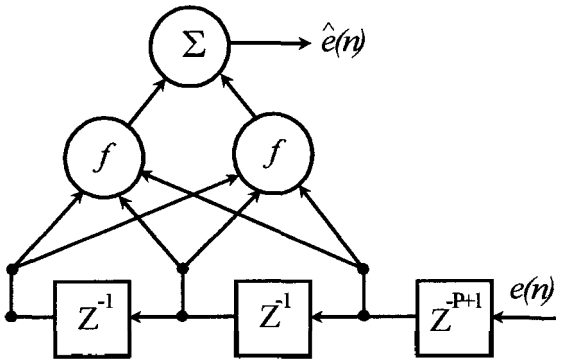

그림 1. 장구간 신경망 예측기
Fig. 1. Long-term neural network predictor.

나타낼 수 있다.

$$\hat{e}(n)=\sum_{j=1}^{2} w_j f\Big(\sum_{i=-1}^{1} v_{ij} e(n-P-i)\Big)=F(\boldsymbol{W},\ \boldsymbol{V},\ \boldsymbol{e}(n)) \tag{1}$$

여기서 w_j는 출력 계층 (output layer)에서의 가중치를 나타내며 v_{ij}는 입력 계층 (input layer)에서의 가중치를 나타낸다. $f(.)$는 전술한 바와 같이 쌍곡선 탄젠트 함수를 나타내며, P는 여기 신호의 주기성을 반영하는 피치 주기를 나타낸다. 선형 장구간 예측기에서는 전영역 탐색법 (full search)에 의해 P를 변동시키며 예측 오차가 최소화되는 예측 계수와 피치 주기 P를 함께 구한다[2]. 신경망 예측기에 이와 동일한 방법을 적용하는 경우, 모든 P 값에 대해 가중치를 학습해야 하며 이 과정에 많은 시간이 소요된다. 따라서 본 논문에서는 P를 미리 찾고, 주어진 P에 의해 신경망을 학습시키는 방법을 사용하였다. 여기서 P는 음성 신호의 피치 추정에 널리 쓰이는 자기 상관법 (clipped autocorrelation method)[1]을 여기 신호에 적용하여 얻도록 하였다.

주어진 여기 신호에 대해 최적의 가중치 w_j^*, v_{ij}^*는 주어진 분석 구간내에서 예측 오차의 자승이 최소화되도록 얻어진다. 예측 오차가 가중치에 대해 볼록 함수 (convex function) 형태를 갖는다고 가정하면, 최적이 가중치는 아래의 최급강하 (steepest descent) 알고리즘에 의해 얻을 수 있다[11].

$$w_j^{(t+1)}=w_j^{(t)}-\frac{1}{2}\eta\Delta_{w_j}\|e(n)-\hat{e}(n)\|^2 \tag{2}$$

$$v_{ij}^{(t+1)}=v_{ij}^{(t)}-\frac{1}{2}\eta\Delta_{v_{ij}}\|e(n)-\hat{e}(n)\|^2 \tag{3}$$

여기서 η는 학습 이득 (learning gain)으로 0 과 1 사이의

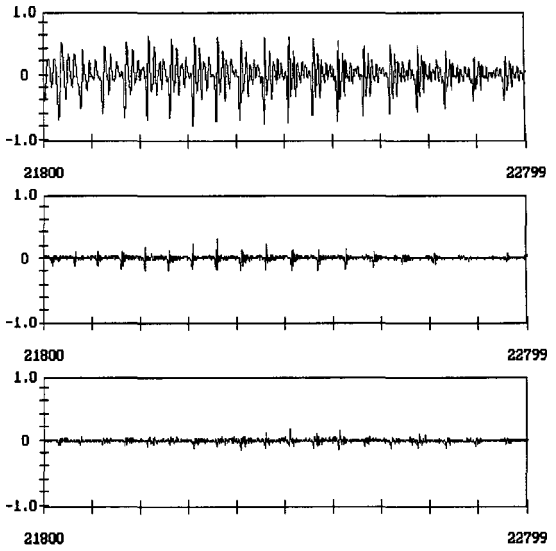

그림 2. 신경망을 이용한 예측후의 잔차신호
(상: 음성신호, 중: 선형예측후 잔차신호, 하: 신경망을 이용한 비선형 예측후 잔차신호)
Fig. 2. Residuals after neural network prediction.
(top: Speech signals, mid: Residuals after linear prediction, bottom: Residuals after nonlinear prediction)

값을 갖는다. 그림 2에 신경망 장구간 예측기를 사용한 결과가 제시되었다. 여기에 제시된 결과는 그림 1에 제시된 것과 동일한 구조의 신경망 예측기를 여성 화자가 발성한 음성에 적용했을 때 얻어진 것이다. 그림 2의 중간에 제시된 선형 예측후의 잔차신호는 펄스 성분이 많이 포함되어 있으나, 이 신호를 비선형 예측기를 통과시켜 얻은 잔차신호는 하단에서 보는 바와 같이 펄스 성분들이 상당수 소멸되어 있음을 알 수 있다.

III. 벡터 양자화기와 결합된 신경 회로망 예측기[4]

앞 장에서 살펴본 신경 회로망 예측기는 여기 신호의 펄스 성분을 효과적으로 제거할 수 있으나, 매 분석 구간마다 가중치의 학습이 이루어져야 하므로 선형 예측기와 비교하여 계산 시간이 증가된다는 단점이 있다. 또한 음성 부호화와 같은 응용 분야에 적용되는 경우, 가중치를 무한대의 정밀도로 표현할 수 없으므로 제한된 비트수의 표현에 따른 근사화 오차 (round-off error)의 영향을 고려해야 한다. Thyssen 등의 실험에 의하면 충분한 양의 학습 데이터로부터 신경망 가중치를 구하고, 여기서 얻어진 가중치 값들을 벡터 양자화에 의해 표현하는 경우,

표 1. 선형/비선형 예측기의 할당 비트수에 따른 예측 이득
Table 1. Prediction gain according to bit-allocations for linear/nonlinear predictor.

Predictor의 갯수(비트수)	예측 이득 (Prediction Gain)	
	신경망 (비선형) 예측기	선형 예측기
32(5)	5.98	4.03
64(6)	6.34	4.62

예측기의 성능이 크게 떨어진다고 보고되었다[4]. 이는 LBG 알고리즘에서 각 벡터 cell의 대표값으로 표현되는 중심 벡터 (central vector)가 단순히 벡터의 값만을 표현하는데 이용한다면 최적의 벡터로 간주할 수 있으나, 신경망의 가중치에서와 같이 비선형 함수의 가중치로 사용되는 경우에는 최적의 벡터로 간주될 수 없기 때문이다.

Thyssen 등에 의해 제안된 비선형 예측 벡터양자화 (NLPVQ: Nonlinear predictive VQ) 기법[4]은 이와 같은 신경 회로망 예측기의 단점을 극복하기 위한 것으로, 신경망의 가중치를 몇 개의 대표 가중치로 제한하여 표현한다. 주어진 여기 신호에 대한 최적의 가중치 벡터는 아래 식과 같이 분석 구간내의 예측 오차 자승합이 최소화되도록 선택된다.

$$\{ W^*, V^* \} = \arg\min_{\{ W, V \} \in K} \sum_{n=1}^{N} \| e(n) - F(W, V, e(n) \|^2 \quad (4)$$

여기서 K는 대표 가중치 벡터들로 구성된 코드북을 나타낸다. 코드북을 구성하는 방법은 아래와 같이 요약할 수 있다.

a) 초기 코드북 $K^{(0)}$ 설정
b) 학습 데이터에 포함에 모든 여기 신호에 대해 식 (4)로 주어지는 최적의 가중치 벡터들을 선택함
c) 동일한 가중치 벡터들이 선택된 여기 신호들에 대해 역전파 알고리즘을 적용, 가중치 벡터들을 갱신함.
d) 갱신된 가중치 벡터들로 코드북 $K^{(n)}$을 재구성
e) 갱신된 코드북으로 예측 오차 계산. 예측 오차가 수렴하면 중지, 그렇지 않으면 b)-d)를 반복 수행.

여기서 일반적인 신경망의 학습과 다른 점은 최적의 가중치를 얻기 위한 역전파 학습 알고리즘이 동일한 가중치 벡터로 분류된 학습 데이터들에 대해서만 이루어진다는 점이다. 따라서 신경망의 가중치 w_j, v_{ij} 는 아래의 식으로 갱신된다.

$$w_j^{(t+1)} = w_j^{(t)} - \frac{1}{2} \eta \sum_{e(n) \in k} \sum_{n=1}^{N} \Delta_{w_j} \| e(n) - \hat{e}(n) \|^2 \quad (5)$$

$$v_{ij}^{(t+1)} = v_{ij}^{(t)} - \frac{1}{2} \eta \sum_{e(n) \in k} \sum_{n=1}^{N} \Delta_{v_{ij}} \| e(n) - \hat{e}(n) \|^2 \quad (6)$$

여기서 k는 주어진 여기 신호 $e(n)$가 속하는 클래스 (class) 즉, 가중치 벡터의 코드 인덱스를 나타낸다.

이와 같은 과정을 통해 구성된 NLPVQ의 예측 성능이 표 1에 요약되었다. 여기에 제시된 값은 남성, 여성 각각 2명의 화자로부터 수집된 약 10,000개의 프레임 (frame)으로부터 얻어진 결과이다. 선형 예측기의 경우는 일반적인 음성 코덱에 널리 쓰이는 레빈슨-더빈 (Levinson-Durbin) 알고리즘[1]으로부터 선형 예측 계수를 구하고, 이를 선스펙트럼쌍 (LSP; Line Spectrum Pair)로 변환한 후 LBG 알고리즘을 적용하여 유한 개의 벡터로 나타내었다. 장구간 예측의 예측 차수 (order)는 3차로 설정했으며, 피치값은 두 예측기에 있어서 앞장에서 제시한 자기상관 기법으로 추정하였다. 표에 제시되었듯이, 유한 개의 코드로 표현하는 경우에도 예측 이득면에서 신경망 예측기가 성능 우위를 보임을 알 수 있다.

IV. 신경망 예측기의 양자화 잡음에 대한 민감도

일반적인 예측 부호화기의 경우, 부호화기에서는 본래의 신호를 알고 있기 때문에 예측도 본래의 신호를 통해 이루어질 수 있으나 복호화기에서는 복원된 신호, 즉 정보 압축시 발생하는 양자화 잡음이 포함된 신호들로부터 예측이 이루어진다. 예측 부호화기에서 복원 신호와 원 신호 사이에 존재하는 잡음은 예측 오차 신호를 유한 개의 값으로 표현하는데서 발생하는 양자화 잡음이 주가 되며 예측기의 특성과는 무관하다[1]. 따라서 주어진 예측기의 양자화 잡음에 대한 민감도는, 양자화 잡음이 포함된 복원 신호를 입력하였을 때와 잡음이 없는 본래의 신호를 입력하였을 때의 차이를 비교하여 구할 수 있다. 본 장에서는 3장에서 살펴본 신경망 예측기의 민감도를 구하는 방법에 대해 알아보고, 그 값을 선형 예측기와 비교하고자 한다.

$e = [e_1, \ldots, e_N]^T$와 $\hat{e} = [\hat{e}_1, \ldots, \hat{e}_N]^T$를 각각 잡음이 없는 여기 신호와 양자화 잡음이 포함된 여기 신호라 한다면, $D_x^2 = \| e - \hat{e} \|^2$는 분석 구간내의 양자화 잡음의 자승합을 나타낸다. 각각의 경우에 대한 예측값의

자승차를 $D_y^2=\{F(\boldsymbol{W},\boldsymbol{V},e)-F(\boldsymbol{W},\boldsymbol{V},\hat{e})\}^2$으로 나타내면, 예측기의 민감도 ρ는 아래와 같이 정의한다[6].

$$\rho=\frac{D_y^2}{D_x^2} \tag{7}$$

Schwarz 부등식을 이용하여 D_y^2를 나타내면 다음과 같다.

$$\begin{aligned} D_y^2 &= [\sum_{j=1}^{M} w_j\{f(\sum_{i=1}^{N} v_{ij}e_i)-f(\sum_{i=1}^{N} v_{ij}\hat{e}_i)\}]^2 \\ &\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}\{f(\sum_{i=1}^{N} v_{ij}e_j)-f(\sum_{i=1}^{N} v_{ij}\hat{e}_j)\}^2 \end{aligned} \tag{8}$$

윗 식에서, 장구간 예측시의 피치값 P를 생략하였다. 윗식을 평균값 정리 $(\frac{f(b)-f(a)}{b-a}=f'(c),\ c\in[a,b])$를 이용하여 다시 나타내면 다음과 같다.

$$D_y^2\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}\{f'(x)(\sum_{i=1}^{N} v_{ij}e_i-\sum_{i=1}^{N} v_{ij}\hat{e}_i)\}^2 \tag{9}$$

여기서 x는 $\sum_{i=1}^{N} v_{ij}e_i$와 $\sum_{i=1}^{N} v_{ij}\hat{e}_i$ 사이의 임의의 값이다. $f(x)=\tanh(x)=\frac{1-\exp[-\alpha x]}{1+\exp[-\alpha x]}$ 일 때 $f'(x)$의 최대 값은 1 이므로, 윗식은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} D_y^2 &\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}\{f'(x)(\sum_{i=1}^{N} v_{ij}e_i-\sum_{i=1}^{N} v_{ij}\hat{e}_i)\}^2 \\ &\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}\{\sum_{i=1}^{N} v_{ij}(e_i-\hat{e}_i)\}^2 \end{aligned} \tag{10}$$

Schwarz 부등식을 다시 한번 적용하면,

$$\begin{aligned} D_y^2 &\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}(\sum_{i=1}^{N} v_{ij}^2 \sum_{i=1}^{N}(e_j-\hat{e}_j)^2) \\ &\le \sum_{j=1}^{M} w_j^2 \sum_{j=1}^{M}\sum_{i=1}^{N} v_{ij}^2 \sum_{i=1}^{N}(e_j-\hat{e}_j)^2 = \|\boldsymbol{W}\|^2\|\boldsymbol{V}\|^2 D_x^2 \end{aligned} \tag{11}$$

여기서 $\|.\|^2$은 Euclidean norm 나타낸다. 식 (11)에서 신경망 예측기의 민감도 ρ는 다음과 같은 조건을 만족함을 알 수 있다.

$$\rho \le \|\boldsymbol{W}\|^2\|\boldsymbol{V}\|^2 \tag{12}$$

선형 예측기의 경우에도 이와 유사한 과정을 통해 민감도를 구하면 예측 계수의 Euclidean norm으로 주어진다[6].

실험적으로 얻어진 신경망 예측기의 민감도를 살펴보았을 때, 값의 범위가 무척 크며 (신경망: $0.62\le\rho\le384$, 선형: $0.14\le\rho\le7.03$), 평균값에 있어서도 선형 예측기의 민감도와 비교하여 약 80배 정도 큰 값을 가짐을 알 수 있었다. 이와 같은 결과는 신경망을 사용한 예측기가 선형 예측기와 비교하여 양자화 잡음에 매우 민감하게 반응한다는 것을 나타내는 것으로, 예측기 자체의 성능이 우수하여도 양자화 과정이 포함된 실제 음성 코덱에서는 선형 예측기를 사용한 경우보다 성능이 저하될 수 있음을 의미한다고 볼 수 있다.

V. 민감도가 고려된 신경망 예측기

본 장에서는 앞장에서 살펴본 기존의 신경망 예측기 문제점을 해결할 수 있는 새로운 기법을 제안하였다. 예측 오차와 민감도를 함께 고려한 신경망 예측기를 설계하기 위하여, 본 논문에서는 신경망의 학습 방법인 역전파 알고리즘을 수정하였다.

예측 오차를 감소시키면서 동시에 민감도를 작게 하기 위하여, 아래와 같은 제한 조건을 갖는 최소화 문제 (minimization problem)를 먼저 정의하였다.

$$\begin{aligned} &\text{minimize} && \{e(n)-\sum_{j=1}^{M} w_j f(\sum_{i=1}^{N} v_{ij}e(n-j))\}^2 \\ &\text{subject to} && \|\boldsymbol{W}\|^2\|\boldsymbol{V}\|^2\le\alpha \end{aligned} \tag{13}$$

여기서 α는 민감도의 최대 허용치이다. 위 문제를 해결하기 위하여, 다음과 같은 라그랑제 (Lagrange) 함수를 정의하였다.

$$\begin{aligned} &L(\boldsymbol{W},\boldsymbol{V},\lambda)= \\ &\{e(n)-\sum_{j=1}^{M} w_j f(\sum_{i=1}^{N} v_{ij}e(n-j))\}^2+\lambda(\alpha-\|\boldsymbol{W}\|^2\|\boldsymbol{V}\|^2) \end{aligned} \tag{14}$$

여기서 λ는 라그랑지 곱수 (Lagrange multiplier)를 나타낸다. 식 (13)을 만족하는 최적의 해 $\{\boldsymbol{W}^*,\boldsymbol{V}^*,\lambda\}$는 $L(\boldsymbol{W},\boldsymbol{V},\lambda)$를 최소화 시킨다. 이러한 최적해는 역전파 알고리즘에서와 같이 반복적인 방법 (iterative method)에 의해 구할 수 있다. 그러나 위의 문제는 부등식의 제한 조건을 포함하고 있으므로, 최적해를 구하는 방법을 새로이 구성할 필요가 있다.

일반적으로 부등식의 제한 조건이 포함된 최소화 문제를 해결하기 위해서는 쿤-터커 조건을 만족하는 함수의 근을 구해야 한다[11]. 쿤-터커 부등식 조건을 만족하는 함수의 근은 아래와 같은 반복 추정 기법에 의해 구할 수 있다.

$$\begin{aligned} v_{ij}^{(l+1)} &= v_{ij}^{(l)}-\frac{1}{2}\eta\Delta_{v_{ij}}L(\boldsymbol{W},\boldsymbol{V},\lambda) \\ &= v_{ij}^{(l)}+\eta w_j e(n)f'(z_j(n))e(n-j)+\eta\lambda^{(l)}v_{ij}^{(l)}\|\boldsymbol{W}\|^2 \end{aligned} \tag{15}$$

$$w_j^{(t+1)} = w_j^{(t)} - \frac{1}{2}\eta\Delta_{w_j}L(W, V, \lambda) = w_j^{(t)} + \eta\varepsilon(n)y_k(n) + \eta\lambda^{(t)}w_j^{(t)}\| V \|^2 \quad (16)$$

여기서 $\varepsilon(n)$, $z_j(n)$ $y_k(n)$은 각각 아래와 같다.

$$\varepsilon(n) = e(n) - \sum_{j=1}^{M} w_j f\left(\sum_{i=1}^{N} v_{ij}e(n-j)\right) \quad (17)$$

$$z_j(n) = \sum_{i=1}^{N} v_{ij}e(n-i) \quad (18)$$

$$y_k(n) = f\left(\sum_{i=1}^{N} v_{ik}e(n-i)\right) \quad (19)$$

라그랑지 곱수 λ에 대한 갱신식은 아래와 같다.

$$\lambda^{(t+1)} = \lambda^{(t)} + \eta(\alpha - \| W^{(t)}\|^2\| V^{(t)}\|^2), \quad \text{if } (\alpha - \| W^{(t)}\|^2\| V^{(t)}\|^2) < 0$$
$$\lambda^{(t)}, \quad \text{if } (\alpha - \| W^{(t)}\|^2\| V^{(t)}\|^2) \geq 0 \quad (20)$$

위의 식으로부터, 라그랑지 곱수는 현재의 신경망 가중치값들이 주어진 제한 조건 $\| W\|^2\| V\|^2 \leq \alpha$를 만족하는 가에 따라 갱신됨을 알 수 있다. 최종적인 해 $\{W^*, V^*, \lambda\}$는 $\Delta L(W, V, \lambda)$가 임계치보다 작은 경우에 얻어진다.

VI. 모의 실험과 결과

앞장에서 제안된 제한조건을 갖는 신경망 학습 기법은 본래의 역전파 학습 알고리즘과는 다른 가중치를 나타낼 수 있으므로, 제안된 기법에 의해 학습된 신경망의 성능을 일반적인 역전파 학습 알고리즘에 의해 학습된 신경망과 비교할 필요가 있다. 본 논문에서는 두 신경망 예측기간의 성능 비교 척도로서 2, 3장에서 도입했던 예측 이득(prediction gain)을 사용하였다.

실험에 사용된 음성 데이터는 여성, 남성 각각 2명의 화자로부터 녹음되었으며, 이를 저전송률 음성 부호화기의 표본화 주파수인 8 KHz로 샘플하고 16 bits로 양자화하였다. 음성 데이터는 배경 잡음이 없는 비교적 조용한 환경에서 녹음되었으며, 따라서 배경 잡음으로 인한 두 예측기간의 성능 영향은 고려하지 않았다.

샘플된 음성 신호에서 수작업에 의해 묵음 구간을 제거하였으며, 장구간 예측기의 효과가 비교적 높게 나타나는 유성음 구간만을 추출하여 최종적인 실험 데이터로 이용하였다. 선형 예측 후의 잔차 신호를 얻기 위하여, 분석 구간마다 10차 LPC 계수를 구하고 음성 신호를 역필터에 통과시켰다. LPC 분석에는 해밍 창함수를 이용하였고, 창함수의 길이는 2.4 kbps의 전송률을 갖는 MELP (Mixed Excitation Linear Predictive) 부호화기[12]에서와 마찬가지로 22.5 msec로 설정하였다. 또한 음성 신호에 포함된 직류 바이어스의 영향을 제거하기 위해 LPC 분석전에 음성 신호를 0.95의 필터 계수를 갖는 전강조(pre-emphasis) 필터에 통과시켰다.

성능 비교를 위해 사용된 음성 데이터는 약 10,000개의 프레임으로서, 3장에서 살펴본 NLPVQ 코드북을 각각의 경우 (제한조건이 포함된 역전파 알고리즘, 포함되지 않은 역전파 알고리즘)에 대해 구성하고, 평균 예측 이득을 구하였다.

제한 조건에 따른 예측 이득을 알아보기 위해 식 (13)의 α를 1.0부터 5.0까지 변화시키고 각 경우에 따른 예측 이득을 계산하였다. 실험시의 가중치 벡터의 개수는 64개로 설정하였으며, 이 경우의 결과가 표 2에 제시되어 있다. $\alpha = 3.0$에서는 예측 이득이 제한 조건이 없는 역전파 알고리즘을 사용한 경우와 비교하여 0.14 dB의 차이만을 나타내므로, 예측기로서의 성능이 크게 저하되지 않음을 알 수 있다. 그러나 $\alpha < 3.0$에서는 예측 이득이 1.0 dB 내외로 떨어지는 등 급격한 성능 저하가 일어난다. 이는 α가 1에 근접하는 경우 제한 조건인 $\| W\|^2\| V\|^2 \leq \alpha$를 만족하는 W, V는 유클리디언 정규값 (Euclidean norm)의 값이 매우 작게 되는 것에 그 원인이 있다. 신경망의 가중치가 매우 작은 값이 되면 예측된 값이 0에 근접한 값이 되며, 결과적으로 예측 오차 신호는 본래의 신호와 거의 유사한 값을 갖기 때문이다. 이러한 경우 예측 이득은 0에 가까운 값을 갖으며, 본 실험에 있어서도 α가 작을수록 예측 이득이 0에 가깝게 됨을 알 수 있었다.

앞장에서 살펴본 선형 예측기의 민감도는 $0.14 \leq \rho \leq 7.03$ 범위의 값임을 알 수 있었다. 따라서 예측 이득면에서 큰 차이를 보이지 않는 $\alpha = 3.0$ 정도의 값으로 최대 허용 민감도를 설정하면, 실제적으로 사용하는데 있어서 큰 문제는 없을 것으로 보인다. 제한 조건이 포함된 신경망 예측기의 경우에도 예측 후 잔차 신호는 그림 2에서 제시한 것과 동일하게 펄스 성분이 거의 제거됨을 알 수 있었으며, 이는 제한 조건이 포함된 신경망 예측기도 여기 신호의 비선형성을 비교적 잘 모델링하는 것이라 말할 수 있다.

실제적인 음성 코덱에 적용시, 신경망 예측기의 민감도는 예측 후 잔차 신호의 양자화 비트수를 결정하는데 사용될 수 있다. 예로서 예측기가 비교적 큰 민감도를 갖는 경우에는 양자화 오차의 영향을 민감하게 받는다는

표 2. 제한조건에 따른 신경망 예측기의 예측 이득
Table 2. Prediction gain of the neural network predictor according to the constraints.

허용 민감도(= α)	예측 이득 (dB)
0.0 (no contraint)	6.34
1.0	1.36
2.0	5.16
3.0	6.20
4.0	6.22
5.0	6.28
Linear	4.62

의미로 해석할 수 있으므로 비교적 많은 비트를 할당해야 하며, 민감도가 적은 경우에는 적은 비트수를 할당할 수 있다. 이는 부호화기에서 비트수의 할당이 분산과 같은 입력 신호의 특성뿐만 아니고 예측기의 특성도 함께 고려해야 함을 의미한다.

제안된 기법을 음성 코덱에 적용시 고려되어야 할 또 한가지 문제점은 비선형 예측 후의 잔차 신호를 어떻게 압축하는가 하는 것이다. CELP형 코더에서와 같이 벡터 코드 형태로 표현하는 방법을 생각할 수 있겠으나, 선형 장구간 예측 후와 비선형 장구간 예측 후에 나타나는 신호는 서로 다른 특성을 가지고 있을 수 있으므로 신호의 특성에 적합한 부호화 방법을 적용하는 것이 부호화 효율을 높일 수 있을 것으로 생각된다. 실험적으로 얻어진 비선형 장구간 예측 후의 잔차 신호는 선형 예측시와 비교하여 크기가 작고, 비교적 긴 분석 구간에서도 펄스 성분이 상당수 사라지므로 기존의 CELP형 부호화기와 비교하여 보다 적은 비트수로 잔차 신호를 표현할 수 있을 것으로 예상된다.

VII. 결 론

본 논문에서는 선형 예측 후에 얻어지는 잔차 신호를 비선형 필터의 하나인 신경망에 의해 예측하는 기법을 제시하고, 예측기의 성능을 향상시키기 위한 새로운 학습 알고리즘을 제안하였다. 신경망 예측기에서의 양자화 오차에 대한 영향을 알아보기 위해 주어진 신경망 예측기의 구조에서 민감도를 구하고, 이 민감도가 허용된 값을 초과하지 않도록 하는 제한 조건을 갖는 신경망 학습 알고리즘을 제안하였다.

제안된 기법의 성능을 분석하기 위해 몇 개의 음성 데이터를 수집하고 선형 예측 잔차 신호를 구성한 다음, 학습을 통해 얻어진 신경망 예측기의 예측 이득을 살펴보았다. 실험적인 결과로부터 허용되는 민감도가 매우 작은 값이 아닌 이상 예측기의 성능은 크게 저하되지 않았으며, 제한 조건이 없는 신경망 예측기와 유사한 예측 이득을 얻을 수 있었다. 이는 제안된 학습 방법에 의해 얻어진 신경망 예측기가 양자와 오차에 대해 강인한 성질을 가지면서, 동시에 비선형 예측기의 장점을 수용한다고 볼 수 있다.

향후 연구 과제로는 예측 후 잔차 신호의 특성을 분석하여 적합한 압축 방법을 모색하고, 비선형 예측기가 포함된 음성 부호화기를 설계하는 것을 들 수 있겠다.

참 고 문 헌

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of speech signals*, Prentice-Hall, 1987.
2. M. Schroeder and B. Atal, "Coded-excited linear prediction (CELP): high-quality speech at very low bit rates," *Proc. ICASSP '85*, 937-940, 1995.
3. J. Thyssen, H. Nielsen and S. D. Hansen, "Non-linear short-term prediction in speech coding," *Proc. ICASSP '94*, 1, 185-188, 1994.
4. J. Thyssen, H. Nielsen and S. D. Hansen, "Quantization of non-linear predictors in speech coding," *Proc. ICASSP '95*, 1, 265-268, 1995.
5. L. Wu, M. Niranjan, and F. Fallside, "Nonlinear predictive vector quantization with recurrent neural nets," *Proc. IEEE-SP Workshop on Neural Networks for signal Processing*, 372-381, 1993.
6. L. Wu, M. Niranjan, and F. Fallside, "Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding," *IEEE Trans. on Speech and Audio Processing*, 2, Issue 4, 482-489, 1994.
7. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, John Wiley & Sons, 1980
8. E. Mumolo, Alberto Carini and Diego Francescato, "ADPCM with non linear predictors," *Proc. EUSIPCO '94*, I, 387-390, 1994.
9. F. Díaz-de-María and A. R. Figueiras-Vidal, "Nonlinear prediction for speech coding using radial basis function," *Proc. ICASSP '95*, 788-791, 1995.
10. R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP megazine*, 4-22, April, 1987.
11. A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*, John Wiley & Sons, 1993.
12. A. V. McCree and T. P. Barnwell III, "A Mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. on Speech and Audio Processing*, 3 (4), 242-250, 1995.

저자 약력

● 이 기 승 (Ki-Seung Lee)

1987년 3월~1991년 2월: 연세대학교 전자공학과 (공학사)
1991년 3월~1993년 2월: 연세대학교 대학원 전자공학과 (공학석사)
1993년 3월~1997년 2월: 연세대학교 대학원 전자공학과 (공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임 연구원
1997년 10월~1999년 8월: AT&T Shannon Lab NJ, USA, Consultant
1999년 9월~2000년 9월: AT&T Shannon Lab NJ, USA, Senior Technical Staff Member
2000년 11월~2001년 8월: 삼성종합기술원, 전문연구원
2001년 9월~현재: 건국대학교 정보통신 대학 전자 공학과, 조교수
※ 주관심분야: 음성 합성, 운율 제어, 음성 변환, 초저전송률 음성 부호화기 등.