선형예측계수를 사용한 신경회로망에 의한 잡음량의 인식

최재승*

*신라대학교

Recognition of Noise Quantity by Neural Network using Linear Predictive Coefficient

Jae-seung Choi*

*Silla University

E-mail : jschoi@silla.ac.kr

요 약

잡음환경 하의 회화에서 잡음량을 줄이고 신호처리 시스템의 성능을 향상시키기 위해서는 잡음량에 따라서 적응적으로 처리되는 신호처리 시스템이 필요하다. 따라서 본 논문에서는 선형예측계수를 사용하여 잡음량을 인식하는 방법을 제안하며, 본 잡음량 인식은 다양한 배경잡음에 의하여 열화된 3종류의 음성이 신경회로망에 의하여 학습되어진다. 본 실험에서는 Aurora2 데이터베이스를 사용하여 여러 잡음에 대하여 평균적으로 약 97.6% 이상의 양호한 인식결과를 확인할 수 있었다.

ABSTRACT

In order to reduce the noise quantity in a conversation under the noisy environment, it is necessary for the signal processing system to process adaptively according to the noise quantity in order to enhance the performance. Therefore this paper presents a recognition method for noise quantity by linear predictive coefficient using a three layered neural network, which is trained using three kinds of speech that is degraded by various background noises. In the experiment, the average values of the recognition results were 97.6% or more for various noises using Aurora2 database.

키워드

Linear predictive coefficient, recognition rate, noise quantity, neural network

Ⅰ. 서 론

근년 음성인식 기술은 음성정보처리 기술의 발달과 더불어 다양한 분야에서 실용화가 진행되고 있으며, 이에 대한 연구가 활발히 진행되고 있다. 그러나 음성인식 기술을 상업적으로 적용하기 위해서는 여러 가지 기술적인 문제를 해결해야 한다. 이러한 문제들 중에서 가장 중요한 요소는 음성에 부가되는 배경잡음의 영향을 줄이는 일이다[1, 2]. 이러한 배경잡음은 인식 대상 음성에 부가되어 음성인식 성능을 크게 저하시키게 된다.

최근에 신경회로망(Neural Network, NN)은 식별에 있어서 상당히 효과적인 능력이 있으며, 음성 및 문자의 식별에 대해서도 많은 성과를 올리고 있다[2, 3]. 또한 음성 중에서 잡음을 경감하기 위해서는 잡음의 강도에 따라서 각각 적당한 처리를 할 필요가 있다. 즉, 잡음의 크기를 인식하는 것이 상당히 중요하다.

본 논문에서는, 배경잡음의 영향을 줄여서 음성인식 시스템의 성능을 향상시키고 다양한 음성인식기의 입력으로 사용하기 위하여, 선형예측분석에 의한 선형예측계수를 신경회로망의 입력으로 한 시스템을 구축하고자 한다. 본 논문에서 사용하는 신경회로망의 입력데이터로는 각각의 프레임의 데이터를 사용하여 학습시키며, 신경회로망의 학습조건 및 학습방법 등을 바꾸어 음성 중의 잡음량을 인식하여 이러한 잡음을 경감하는 것을 목적으로 한 연구를 진행한다.

본 연구의 목적을 달성하기 위하여 본 논문에서

는 잡음과 음성신호의 특징을 가진 선형예측계수(Linear Predictive Coefficient, LPC)를 신경회로망의 입력으로 하여 3종류의 잡음량을 인식하는 방법을 제안한다.

II. 음성신호의 선형예측분석

음성신호의 표본값 사이에는 커다란 상관관계가 있으며 음성의 특징 추출을 위하여 이것을 이용한 예측부호화가 실시되어진다[4]. 이러한 예측의 개념을 일반화하여 다음 식과 같이 음성파형의 연속된 $p+1$개의 표본값 사이에 높은 선형예측성이 있다고 가정한다.

$$\hat{x_n} = a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_p x_{n-p} + e_n \cdots (1)$$

따라서 음성신호의 현재 값은 p개의 과거의 값 $x_{n-1}, x_{n-2}, \cdots, x_{n-p}$로부터 예측된다. 여기에서 $a_{i(i=1,\cdots,p)}$는 선형예측계수이며, e_n은 식 (2)와 같이 실제 입력된 값과 예측된 값과의 차이를 나타내는 선형예측오차이다.

$$e_n = x_n - \hat{x}_n \cdots\cdots (2)$$

식 (1)로부터 선형예측오차 e_n의 2승 평균값을 계산하여 장시간 평균을 구한다. 이 때, 시간평균 조작은 선형한 조작 방법을 이용하여 평균적인 각항 별로 평균한다. 또한 선형예측계수는 시간 불변임으로 다음 식과 같이 구해진다.

$$\begin{aligned}\overline{e_n^2} &= v_{00} + a_1 v_{01} + a_2 v_{02} + \cdots + a_p v_{0p} \\ &\quad + a_1 v_{10} + a_1^2 v_{11} + a_1 a_2 v_{12} + \cdots + a_1 a_p v_{1p} \\ &\quad \cdots\cdots\cdots\cdots \\ &\quad + a_p v_{p0} + a_p a_1 v_{p1} + a_p a_2 v_{p2} + \cdots \\ &\quad \cdots + a_p a_{p-1} v_{pp-1} + a_p^2 v_{pp} \cdots\cdots\cdots\cdots\cdots (3)\end{aligned}$$

여기에서 $v_{ij} = \sum_n x_{n-i} \cdot x_{n-j}$으로 파형 $\{\hat{x_n}\}$의 상관계수이다. 위의 식을 최소로 하는 $\{a_i\}$의 조건으로는, 각 a_i에 의한 편미분을 0으로 했을 때 신호가 정상임으로, $v_{ij} = v_{|i-j|}$를 사용하여 다음의 p개의 식이 구해진다.

$$\begin{aligned}&2(a_1 v_0 + a_2 v_1 + \cdots + a_p v_{p-1} + v_1) = 0 \\ &2(a_1 v_1 + a_2 v_0 + \cdots + a_p v_{p-2} + v_2) = 0 \\ &\qquad\cdots\cdots \\ &2(a_1 v_{p-1} + a_2 v_{p-2} + \cdots + a_p v_0 + v_p) = 0 \cdots (4)\end{aligned}$$

구하는 해는 이 p식을 연립하는 것이며, 이 결과에 의하여 극치가 구해져, 이것이 유일한 해이면 구하는 최소치이다. 위의 식을 정리하면 다음식과 같은 연립 p원 1차방정식이 된다.

$$\begin{vmatrix} v_0 & v_1 & v_2 \cdots v_{p-1} \\ v_1 & v_0 & v_1 \cdots v_{p-2} \\ v_2 & v_1 & v_0 \cdots v_{p-3} \\ \cdots & \cdots\cdots\cdots & \\ v_{p-1} & v_{p-2} & v_{p-3} \cdots v_0 \end{vmatrix} \begin{vmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ p \end{vmatrix} = - \begin{vmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_p \end{vmatrix} \cdots (5)$$

이와 같이 하여 구해진 선형예측계수는 분석의 대상인 일련의 데이터를 전극 모델에 의하여 생성하였을 때의 시스템의 요소가 된다. 따라서 이 계수로서 생성 모델의 정보가 추출되어, 이것들을 부호화함으로서 고능률 부호화가 가능하다.

III. 신경회로망을 사용한 선형예측계수에 의한 잡음량 인식

II장에서 기술한 방법으로, 음성신호의 표본값을 선형예측 분석하여, 본 실험에서는 10차의 선형예측계수를 구한다. 이렇게 함으로써, 원래의 표본값은 10차의 선형예측계수와 잔차신호로 완전히 복원가능하다. 분석대상으로 하는 음성신호의 표본값에 잡음이 중첩된 경우, 잡음은 선형예측계수와 잔차신호의 모두에 영향을 미치지만, 본 실험에서는 잔차신호는 조작을 하지 않고 선형예측계수 만을 사용한다. 이 값의 차이로부터 잡음량을 인식 하는 것을 목적으로 하며, 이 목적을 위한 방법으로 신경회로망을 사용한다. 본 논문에서 사용한 신경회로망의 학습법은 Rumelhart[5]에 의해서 제안된 2승 오차최소화의 학습을 다층 네트워크 전체의 학습에 확장시킨 방법으로 오차역전파학습법으로 불리운다. 이 학습법은 입력값이 주어졌을 때 교사신호와 출력값의 오차를 최소로 하며 신경세포 사이의 결합계수를 조절하는 방법이다. 본 실험에서는, 잡음이 없는 음성(SNRin(Input Signal-to-Noise Ratio)=∞), 잡음이 적은 음성(SNRin=15dB), 잡음이 많은 음성(SNRin=5dB)의 3종류를 인식할 수 있도록, 신경회로망의 출력층의 유닛수를 3으로 하여 학습을 시킨다. 이것으로부터 구해진 결과는 실제의 잡음량과 비교되어 인식율의 형태로 평가한다.

이상과 같은 음성에 의하여 구해진 10차의 선형예측계수는 입력층의 각 유닛에 입력되며, 신경회로망의 교사신호는 (T1) SNRin=∞일 때 (1.0, -1.0, -1.0), (T2) SNRin=15dB 일 때 (-1.0, 1.0, -1.0), (T3) SNRin=5dB 일 때, (-1.0, -1.0, 1.0)으로 한다. 그리고 각각의 유닛의 하중은 -0.005~0.025의 범위 내의 랜덤한 값으로 초기화하며, 학습계수 $\alpha = 0.2$, 가속도계수 $\beta = 0.5$로 한다. 본 실험에서는, 3층 구조의 퍼셉트론(perceptron)형[6]의 신경회로망의 구조인 입력층의 10 유닛, 중간층의 15 유닛, 출력층의 3 유닛으로 구성된 네트워크를 사용하였다. 또한 학습의 횟수를 5000회로 하여 각 음성데이터에 있어서 결합하중의 초기값을 바꾸어서 5회씩 시행한다.

본 실험에서 사용한 음성 데이터는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스[7]이다. 제안한 시스템은 Aurora2 데이터베이스로부터의 테스트 셋 A, B, C의 음성데이터와 테스트 셋 A의 자동차(car noise), 지하철잡음(subway noise),

테스트 셋 C의 도로잡음(street noise) 그리고 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise) 등의 배경잡음을 사용하여 평가하였다. 본 실험에서는 3종류의 입력 신호대잡음비 (SNRin= ∞, 15 dB, 5 dB)와 같이 잡음이 부가된 음성신호를 사용하여 신경회로망을 학습시켰다. Aurora2 데이터베이스를 사용할 경우에 백색잡음, 자동차잡음, 지하철잡음을 Aurora2 데이터베이스의 음성신호에 부가한 후에 신경회로망이 학습되었다.

본 실험에서는 음성데이터에 대, 중, 소 3종류의 잡음 데이터를 각각 부가하여, 각각의 데이터에 대하여 각 프레임 별로 선형예측계수를 구한다. 여기에서 1프레임은 256샘플로 한다. 단, 본 실험에서는 각 음성데이터의 평균전력 R_m을 구하여 각 프레임에서의 평균전력 R_f와의 관계가 다음식과 같이 되도록 프레임의 선형예측계수 만을 사용한다.

$$R_m = \sqrt{\frac{\sum_{i=1}^{M} S_i^2}{M}} \quad \cdots\cdots (6)$$

$$R_f = \sqrt{\frac{\sum_{i=1}^{256} S_i^2 \cdot_f}{256}} \quad \cdots\cdots (7)$$

$$\frac{R_m}{3} > R_f \quad \cdots\cdots (8)$$

단, M은 전샘플수, S_i는 문장 전체에서 구한 잡음이 중첩된 음성신호의 표본값, $S_{i\cdot f}$는 각 프레임 내의 잡음이 중첩된 음성신호를 나타낸다.

학습 문장 사이에서의 패턴의 유사성을 측정하기 위해서 아래의 식과 같이 정의된 선형예측계수의 거리 D를 측정하여, 거리와 신경회로망에 의한 잡음량의 인식율과의 관계를 명확하게 하였다.

$$D_{i\cdot j} = \sqrt{\sum_{k=1}^{10} (I_k - J_k)^2} \quad \cdots\cdots (9)$$

여기에서, $i \cdot j$는 비교하는 잡음의 크기를 나타낸다. 예를 들면, T2(SNRin=15dB)와 T3(SNRin=5dB)의 비교라면, $D_{T2\cdot T3}$가 된다. I_k, J_k는 각각 비교하는 프레임의 선형예측계수의 값이다.

표 1은 선형예측계수 방식에 의한 거리측정결과와 켑스트럼계수 방식에 의한 거리측정결과를 나타내고 있으며, 학습에 사용된 10개의 네트워크에 대한 평균값의 결과이다. 켑스트럼계수 방식은 입력음성을 해밍창을 통과 시킨 후에 켑스트럼변환하여 저역에 해당하는 10개의 켑스트럼계수를 구하는 방식이다. 표의 결과로부터, 1프레임 중의 선형예측계수의 거리 D의 값이 SNRin=15dB(T2)과 SNRin=5dB(T3)의 경우($D_{T2\cdot T3}$), 다른 경우와 비교하면 거리가 극히 밀접하기 때문에, 신경회로망에 의한 인식이 쉽지 않을 거라고 추측되어진다. 또한 켑스트럼거리 방식이 본 논문에서 제안한 선형예측계수 방식보다 거리가 상당히 떨어져 있으므로, 각 데이터간의 중첩이 적고 신경회로망에 의한 인식이 본 방식보다 어느 정도 용이하게 가능하다는 것을 판단할 수 있다. 이러한 방식들에 의한 인식 결과를 제IV장에서 자세히 나타낸다.

표 1. 선형예측계수 및 켑스트럼 거리의 비교

	$D_{T1\cdot T2}$	$D_{T1\cdot T3}$	$D_{T2\cdot T3}$
linear predictive	0.359	0.396	0.049
cepstral	0.635	0.714	0.221

그림 1은 입력으로 하는 선형예측계수의 예를 그래프로 나타낸 것으로(제10프레임의 선형예측계수), 본 실험에서 사용한 프레임의 선형예측계수 뿐만 아니라 거의 전 프레임에 대해서, SNRin=15dB(T2)과 SNRin=5dB(T3)에 대하여 선형예측계수에는 거의 차이 없었다. 그림에서 알 수 있듯이, 선형예측계수의 간격이 상당히 접근해 있음에도 불구하고 제IV장에서 나타내는 인식율 실험과 같이 각 학습데이터를 양호하게 인식할 수 있었다.

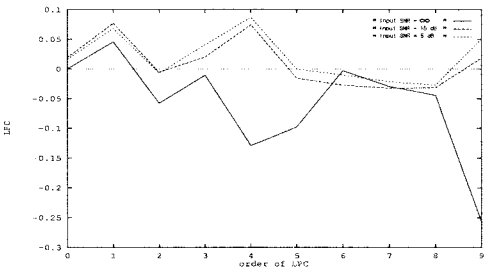

그림 1. 잡음량의 차이에 의한 선형예측계수의 예

IV. 실험 결과

본 논문에서는 출력되어진 학습결과와 학습신호를 비교하여 인식율을 구한다. 인식율 P는 다음식과 같이 정의한다.

$$P = \frac{F_c}{F_s} \times 100 \quad \cdots\cdots (10)$$

여기에서, F_c는 정확하게 인식된 프레임수를 나타내고, F_s는 식 (8)을 만족하는 프레임수를 나타낸다.

표 2와 표 3은 Aurora 2 데이터베이스의 테스트 셋 C로부터 임의적으로 20개의 문장을 선택하여, 각 음성으로부터 구한 선형예측계수를 신경회로망의 입력으로 하여 실험을 실시한 학습 결과를 각 잡음에 대하여 나타낸 평균이다. 표 2는 본 논문에서 제안한 선형예측계수에 의한 잡음량의 인식율을 나타내며, 표 3은 비교를 위하여 켑스트럼

계수 방식에 의한 인식율을 나타낸다.

표 2. 선형예측계수에 의한 각 잡음에 대한 인식율

Type of noise	Recognition rates (%)		
	T1 (∞)	T2 (15 dB)	T3 (5 dB)
white	100.0%	98.7%	100.0%
Car	99.7%	97.4%	100.0%
Subway	98.4%	96.2%	98.8%
Street	96.2%	92.6%	92.6%
Average	**98.6%**	**96.2%**	**97.9%**

표 3. 켑스트럼계수에 의한 각 잡음에 대한 인식율

Type of noise	Recognition rates (%)		
	T1 (∞)	T2 (15 dB)	T3 (5 dB)
white	100.0%	99.1%	100.0%
Car	99.9%	97.8%	100.0%
Subway	99.0%	97.1%	99.5%
Street	97.6%	93.7%	94.9%
Average	**99.1%**	**96.9%**	**98.6%**

표 2의 선형예측계수에 의한 각 잡음에 대한 인식율의 평균값과 표 3의 켑스트럼계수에 의한 각 잡음에 대한 인식율의 평균값을 비교하면, 켑스트럼계수에 의한 방법이 선형예측계수에 의한 인식율보다 약 0.57% 정도의 극히 미세하게 양호하지만, 본 논문에서 제안한 선형예측계수에 의한 인식율은 여러 잡음에 대하여 평균적으로 약 97.6% 이상의 높은 인식결과를 확인할 수 있었다. 또한 본 논문에서 제안한 표 2의 선형예측계수에 의한 3 패턴의 학습신호에 의한 학습결과로부터, SNRin=15dB(T2)의 인식율이 다른 입력(T1 및 T3)보다 약간 인식율이 떨어지는 반면에, SNRin=∞(T1)과 SNRin=5dB(T3)에서 상당히 좋은 인식결과를 볼 수 있었다.

이상의 결과로부터, 제III장의 표 1의 거리측정 결과에서 알 수 있듯이 선형예측계수의 거리가 켑스트럼계수의 거리보다 상당히 밀접해 있음에도 불구하고 선형예측계수에 의한 인식율이 거의 동등하게 인식되었다는 것은 본 논문에서 제안한 방식이 상당히 유효하다는 것을 말할 수 있다.

IV. 결론

본 논문에서는 신경회로망에 의한 3종류의 음성신호의 잡음량을 인식하는 것을 목적으로 하여, 선형예측계수를 입력으로 한 잡음량 인식의 실험을 실시하였다. 이 결과, 신경회로망의 파라미터 및 학습횟수 및 입력데이터의 균형에 따라서 본 논문에서 추구하는 목적을 달성할 수 있었다. 그러나 음성인식 및 음성강조에 응용하기 위해서는, 다음과 같은 내용이 향후의 과제 및 목표라고 생각하며, 추후 더 상세한 연구를 실시하고자 한다. (1) 학습데이터에 정확하고 신속하게 수속하기 위한 중간층의 층수 및 유닛수의 선택, 학습 횟수를 늘인다. (2) 본 실험에서는, 식(8)의 조건에 적합한 프레임 만을 사용하였지만, 역으로 이 조건에 적합하지 않은 프레임을 사용하여 실험한다.

이상과 같이 다양한 잡음이 중첩된 음성신호에 대한 잡음량의 인식을 신경회로망을 통하여 실험적으로 확인하여 본 연구가 음성인식 및 음성신호처리에 효과적으로 응용될 것이라고 생각한다.

참고문헌

[1] S. V. Vaseghi, B. P. Milner: Speech Recognition in Impulsive Noise. International Conference on Acoustics, Speech, and Signal Processing, pp. 437 - 440, vol. 1, 1995.

[2] K. K. Paliwal, "Neural net classifiers for robust speech recognition under noisy environments", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 429-432, April 1990.

[3] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement", IEEE Trans. Speech and Audio Processing, Vol. 3, No. 6, pp. 433-438, 1995.

[4] P.B. Patil: Multilayered network for LPC based speech recognition. IEEE Transactions on Consumer Electronics, Vol. 44, No. 2, pp. 435 - 438, 1998.

[5] D. Rumelhart, "Parallel Distributed Processing, vol. 1 and 2, MIT Press, Cambridge, MA, 1986.

[6] S. K. Pal, S. Mitra, "Multilayer perceptron, fuzzy sets, and classification", IEEE Transaction on Neural Networks, vol. 3, no. 5, pp. 683-697, 1992.

[7] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", in Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 2000.