배경잡음 하에서의 신경회로망에 의한 남성화자 및 여성화자의 성별인식 알고리즘

최재승*

*신라대학교 전자공학과

E-mail : *jschoi@silla.ac.kr

요 약

본 논문에서는 잡음 환경 하에서 남녀 성별인식이 가능한 신경회로망에 의한 화자종속 음성인식 알고리즘을 제안한다. 본 논문에서 제안한 음성인식 알고리즘은 남성화자 및 여성화자를 인식하기 위하여 LPC 켑스트럼 계수를 사용하여 신경회로망에 의하여 학습된다. 본 실험에서는 백색잡음 및 자동차잡음에 대하여 신경회로망의 네크워크에 대한 인식결과를 나타낸다. 인식실험의 결과로부터 백색잡음에 대해서는 최대 96% 이상의 인식률, 자동차잡음에 대해서는 최대 88% 이상의 인식률을 구하였다.

키워드

화자종속 음성인식알고리즘, 신경회로망, LPC 켑스트럼계수, 백색 및 자동차잡음

Ⅰ. 서 론

남녀 성별인식에 사용되는 음성특징 파라미터는 응용분야에 따라서 여러 방법[1]이 제안되고 있다. 음성신호의 파라미터는 음성신호를 어떤 공간의 특징벡터로 사상시키는 것이므로, 이러한 의미에서 본 논문에서는 음성의 발성기관을 LPC (Linear Predictive Coding)[2] 분석하여 구해지는 LPC 켑스트럼 계수로부터 특징벡터의 파라미터를 추출하는 방법을 먼저 제안한다. 그리고 백색잡음 및 자동차잡음이라는 배경잡음 환경 하에서 남성화자 및 여성화자를 인식하기 위한 화자종속 남녀성별인식 알고리즘을 제안한다.

한편 불특정 화자에 대한 음성인식을 할 경우에 각 화자마다 발성기관과 발성습관이 서로 다르기 때문에 특징벡터의 파라미터를 추출하기 어려워 높은 인식률을 구하기가 어렵다. 그러나 본 실험에서는 LPC 켑스트럼 계수가 남성 및 여성화자를 구별할 수 있는 언어정보를 충분히 포함하고 있다고 판단하여, LPC 켑스트럼 계수를 음성인식 파라미터로 사용하여 신경회로망의 학습 알고리즘을 사용하여 백색잡음 및 자동차잡음 중에서도 충분히 남녀의 성별을 인식할 수 있는 실험을 실시한다.

Ⅱ. 제안한 알고리즘

그림 1은 본 논문에서 사용한 남녀성별인식을 위한 3층으로 구성된 퍼셉트론형의 순방향 계층형 신경회로망[3, 4]의 학습과정을 나타낸다. 제안한 신경회로망의 구조는 10, 12개의 입력층 유닛, 20, 30개의 중간층 유닛, 남성화자 및 여성화자를 분류하기 위한 2개의 출력층 유닛을 갖는 네트워크이다. 신경회로망의 학습 계수는 α=0.2, 가속도 계수는 β=0.03으로 하였으며, 최대 학습 횟수는 20,000회로 하였다. 제안한 신경회로망에서는 LPC 켑스트럼 계수(신경회로망의 입력으로 사용)에 대하여 중간층을 20 유닛과 30 유닛의 2종류로 분류하여 총 6개의 네트워크로 구성하여 신경회로망을 학습시킨다.

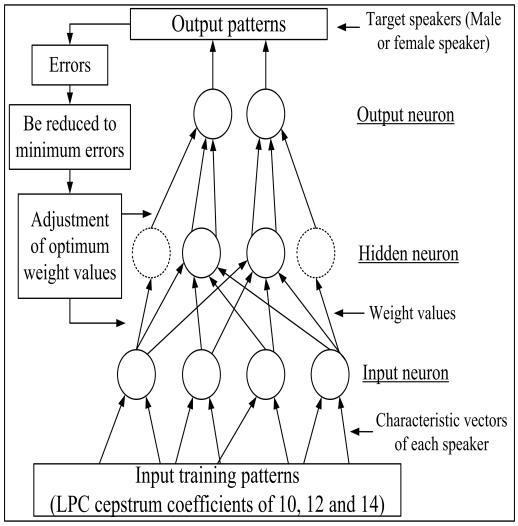

그림 1. 제안한 신경회로망의 구조

그림 2는 본 논문에서 제안한 잡음이 중첩된 음성신호에 대한 화자종속 성별인식 알고리즘의 블록도이다. 백색잡음 및 자동차잡음이 중첩된 음성신호의 한 프레임을 256샘플로 분리한 후에 해밍창을 통과시킨다. 이 후에 LPC 분석과정을 거친 후에 10차, 12차의 LPC 켑스트럼 계수를 추출한다. 추출된 LPC 켑스트럼 계수는 -1.0~+1.0 사이의 값으로 정규화된 후에 이를 신경회로망의 입력층에 입력하여, 오차역전파 학습 알고리즘[3]에 의하여 남성화자 및 여성화자로 인식하도록 학습 과정을 거친다.

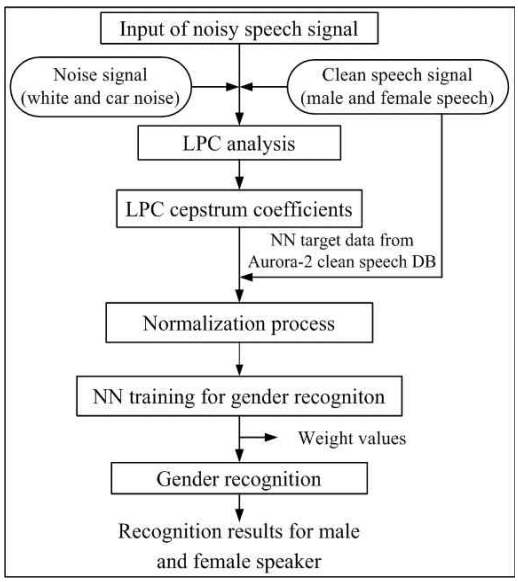

그림 2. 화자종속 인식 알고리즘

III. 실험 결과

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스(Database, DB)[5] 를 사용한다. 본 논문에서 제안한 시스템은 Aurora2 DB로부터의 테스트 셋 A, B, C의 음성데이터와 테스트 셋 A의 자동차(car noise), 그리고 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise) 등의 배경잡음을 사용하여 평가하였다. 본 실험에서는 입력 신호대잡음비 (SNRin(Input Signal-to-Noise Ratio))를 약 -9 dB~-15 dB 사이의 백색잡음이 부가된 음성신호와 SNRin를 약 -8 dB~-16 dB 사이의 자동차잡음이 부가된 음성신호를 사용하여 신경회로망을 학습시켰다. 본 실험에서 사용한 음성데이터는 남성화자의 경우에는 약 50 프레임에서 70 프레임 정도, 여성화자의 경우에는 약 50 프레임에서 85 프레임 정도로 나누워지며, 1 프레임은 256 샘플로 구성된다.

본 실험에서는 10차, 12차의 LPC 켑스트럼 계수를 신경회로망에 입력하였을 경우에 신경회로망의 입력층 3종류 및 중간층 2종류로 구성된 총 6개의 네트워크에 대하여, 백색잡음 및 자동차잡음이 중첩된 음성신호에 대하여 학습에 사용한 남성 및 여성화자에 대한 성별 인식률을 표 1, 표 2에 나타낸다. 표에서 알 수 있듯이, 표 1의 10차의 LPC 켑스트럼 계수에 대해서는 10-20-2 네트워크의 인식률이 양호하였으며, 표 2의 12차의 LPC 켑스트럼 계수에 대해서는 12-30-2 네트워크의 인식률이 양호하였다. 표의 결과로부터 남성화자보다는 여성화자의 인식률이 더 양호하였으며, 또한 자동차잡음보다는 백색잡음의 인식률이 양호하였다. 이러한 결과는 자동차잡음과 같은 유색잡음 혹은 학습 음성데이터의 편중에 의하여 신경회로망의 학습 결과가 약간 상이하다고 볼 수 있다. 따라서 향후의 연구에는 이러한 내용을 개선하기 위한 신경회로망의 학습 조건 및 다양한 학습 데이터를 사용할 필요가 있다고 본다.

표 1 10차의 LPC 켑스트럼에 대한 성별 인식

Speaker	Gender recognition rates[%]			
	10-20-2 network		10-30-2 network	
	White	Car	White	Car
Male	89.38%	86.73%	77.88%	62.83%
Female	100.00%	90.27%	99.12%	92.04%
Average	**94.69%**	**88.50%**	88.50%	77.44%

표 2. 12차의 LPC 켑스트럼에 대한 성별 인식

Speaker	Gender recognition rates[%]			
	12-20-2 network		12-30-2 network	
	White	Car	White	Car
Male	70.80%	61.06%	87.61%	71.68%
Female	98.23%	98.23%	95.58%	92.92%
Average	84.52%	79.65%	**91.60%**	**82.30%**

Ⅳ. 결론

본 논문에서는 저주파수 영역에서의 스펙트럼 구조를 대상으로 한 LPC 분석을 실행하여, 이 적합도를 나타내는 특징 파라미터로부터 음성의 특징벡터를 검출하였다. 따라서 이러한 적절한 음성특징 파라미터를 추출하여 신경회로망에 학습시킴으로써 배경잡음 환경 하에서도 남성화자 및 여성화자를 인식할 수 있는 화자종속 남녀성별인식 알고리즘을 구현하였다.

본 논문에서는 특히 남녀화자의 음성신호에 백색잡음 및 자동차잡음이 중첩된 경우에도 음성인식 실험을 하여, 기존의 방법과 비교하여 본 알고리즘이 효과적인 것을 실험적으로 나타낸다. 또한 SNRin이 낮은 환경 하에서도 특히 백색잡음에 대하여 기존의 방법보다 개선된 결과를 구하였다. 향후 연구 과제로는 좀 더 많은 데이터셋 및 비정상적인 배경잡음에 대하여 독립적인 화자인식이 가능한 알고리즘을 연구할 예정이다.

참고문헌

[1] H. Xu, X. Zhang and L. Jia, "The extraction and simulation of Mel frequency cepstrum speech parameters", 2012 International Conference on Systems and Informatics, pp. 1765-1768, 2012.

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

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors", Nature, Vol. 323, pp. 533-536, 1986.

[4] T. T. Le, J. S. Mason and T. Kitamura, "Characteristics of multi-layer perceptron models in enhancing degraded speech", Proc. ICSLP-94, pp. 1611-1614, 1994.

[5] 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.