남녀성별 분류를 위한 화자종속 음성인식 알고리즘

최재승*

Speaker-dependent Speech Recognition Algorithm for Male and Female Classification

Jae-Seung Choi*

요 약

본 논문에서는 백색잡음 및 자동차잡음 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 총 6개의 신경회로망의 네크워크에 대한 인식결과를 나타낸다. 인식실험의 결과로부터 백색잡음에 대해서는 최대 96% 이상의 인식률, 자동차잡음에 대해서는 최대 88% 이상의 인식률을 구하였다. 마지막으로 본 실험에서는 제안하는 음성인식 알고리즘이 배경잡음 환경 하에서의 기존의 음성인식 알고리즘과 비교하여 본 방식의 알고리즘이 유효하다는 것을 실험으로 확인한다.

ABSTRACT

This paper proposes a speaker-dependent speech recognition algorithm which can classify the gender for male and female speakers in white noise and car noise, using a neural network. The proposed speech recognition algorithm is trained by the neural network to recognize the gender for male and female speakers, using LPC (Linear Predictive Coding) cepstrum coefficients. In the experiment results, the maximal improvement of total speech recognition rate is 96% for white noise and 88% for car noise, respectively, after trained a total of six neural networks. Finally, the proposed speech recognition algorithm is compared with the results of a conventional speech recognition algorithm in the background noisy environment.

키워드

Speaker-dependent speech recognition algorithm, neural network, LPC cepstrum coefficients, white and car noise

Key word

화자종속 음성인식알고리즘, 신경회로망, LPC 켑스트럼계수, 백색 및 자동차잡음

* 정회원 : 신라대학교 전자공학과 교수(주저자, jschoi@silla.ac.kr)

접수일자 : 2012. 08. 28
심사완료일자 : 2012. 10. 08

Open Access http://dx.doi.org/10.6109/jkiice.2013.17.4.775

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

실제 환경에서의 음성인식 성능은 백색잡음과 같은 정상잡음 외에 자동차잡음, 거리잡음, 공항잡음, 전철역잡음, 지하철역잡음 등의 비정상적인 잡음 등에 의하여 급격한 음성인식의 성능열화의 문제점을 가지고 있다. 따라서 이러한 환경 배경잡음에 대하여 음성인식 장치를 실현시키기 위하여 여러 분야에서 연구가 계속적으로 수행되고 있는 실정이다. 현재 대표적인 음성인식 알고리즘으로는 확률적인 방법으로 잘 알려진 은닉 마코프 모델(Hidden Markov Models, HMM) 알고리즘[1], 인간의 신경망을 모델링한 신경회로망(Neural network, NN)[2, 3], 패턴 정합방식인 DTW(dynamic time warping) 방법[4] 등이 있다. 이중에서도 본 논문에서는 신경회로망을 음성인식 모듈에 응용하여, 남성화자 및 여성화자에 대한 음성인식[5-7] 알고리즘을 제안한다.

남녀성별인식에 사용되는 음성특징 파라미터는 응용분야에 따라서 여러 방법[8, 9]이 제안되고 있다. 음성신호의 파라미터는 음성신호를 어떤 공간의 특징벡터로 사상시키는 것이므로, 이러한 의미에서 본 논문에서는 음성의 발성기관을 LPC(Linear Predictive Coding)[9] 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. 또한 피치성분을 가진 RASTA-PLP (Relative Spectral Perceptual Linear Predictive) 계수를 사용하는 기존의 성별분류 방법[5]과 비교하여 본 알고리즘이 효과적인 것을 실험적으로 나타낸다.

Ⅱ. 제안한 다층 퍼셉트론 순방향 신경회로망

그림 1은 본 논문에서 사용한 남녀성별인식을 위한 3층으로 구성된 퍼셉트론형의 순방향 계층형 신경회로망[2, 3]의 학습과정을 나타낸다. 제안한 신경회로망의 구조는 10, 12, 14개의 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 학습횟수는 20,000회로 하였다. 표 1은 본 논문에서 제안한 신경회로망에 사용한 네트워크를 나타낸다. 제안한 신경회로망에서는 LPC 켑스트럼 계수(신경회로망의 입력으로 사용)에 대하여 중간층을 20 유닛과 30 유닛의 2종류로 분류하여 총 6개의 네트워크로 구성하여 신경회로망을 학습시킨다.

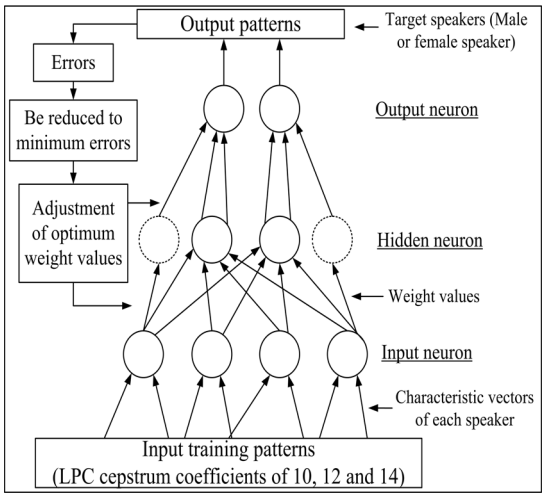

그림 1. 제안한 신경회로망의 구조
Fig. 1 Construction of the proposed neural network

표 1. 제안한 신경회로망의 학습에 사용한 네트워크
Table. 1 Networks of proposed NN training

LPC cepstrum coef.	Networks of proposed NN	
10	10-20-2	10-30-2
12	12-20-2	12-30-2
14	14-20-2	14-30-2

Ⅲ. 화자종속 성별인식 알고리즘

선형예측분석을 음성검출에 이용하는 연구에서는 음성분야에 대한 일반적인 주파수 대역에서 최적 선형예측계수를 구하여, 주기성신호의 비율 등의 정보를 조합한 검출방법의 연구가 보고되고 있다[9]. 이러한 방법들과는 달리 본 논문에서는 저주파수 영역에서의 유성음 특유의 스펙트럼 미세구조를 대상으로 한 LPC 분석을 실행하여, 이 적합도를 나타내는 음성특징 파라미터를 검출한다. 본 논문에서 제안하는 음성의 특징벡터 검출방법은 잡음의 전력 및 주파수 스펙트럼의 변동에 영향을 쉽게 받기 어렵고 종래 방법과 비교하여 비정상적인 고잡음 하에서도 높은 정밀도로 특징벡터의 검출이 가능하다. 따라서 본 논문에서는 대상으로 하는 잡음 환경 하에서의 음성의 특징벡터의 검출 방법으로서 고역과 비교하여 잡음의 영향이 적은 저역에 착목한 분석 방법을 검토한다.

본 실험에서의 LPC 켑스트럼 계수[9]는 10차, 12차, 14차를 사용한다. 그림 2는 신경회로망의 학습에 사용한 남성화자에 대한 10차의 LPC 켑스트럼 계수의 한 프레임(10 프레임)의 모양을 나타내고 있다. 그림 3과 그림 4는 남성화자의 음성신호에 백색잡음 및 자동차잡음이 중첩된 경우에 대한 10차의 LPC 켑스트럼 계수의 변화 모양을 각각 나타낸다. 이상의 그림으로부터 백색잡음 및 자동차잡음이 원음성신호에 중첩됨으로써 남성화자의 원신호의 모양이 변화하는 것을 알 수 있다. 그러나 본 논문에서 제안하는 저주파 영역에서 추출하는 LPC 켑스트럼 계수를 사용함으로써 2장에서 기술한 신경회로망의 학습 알고리즘에 의하여 남성 및 여성화자를 충분히 인식 가능하다는 것을 4장의 인식 실험에서 확인할 수 있다.

그림 5는 본 논문에서 제안한 잡음이 중첩된 음성신호에 대한 화자종속 성별인식 알고리즘의 블록도이다. 백색잡음 및 자동차잡음이 중첩된 음성신호의 한 프레임을 256샘플로 분리한 후에 해밍창을 통과시킨다. 이후에 LPC 분석과정을 거친 후에 10차, 12차, 14차의 LPC 켑스트럼 계수를 추출한다. 추출된 LPC 켑스트럼 계수는 -1.0～+1.0 사이의 값으로 정규화된 후에 이를 신경회로망의 입력층에 입력하여, 오차역전파 학습 알고리즘[2]에 의하여 남성화자 및 여성화자로 인식하도록 학습과정을 거친다.

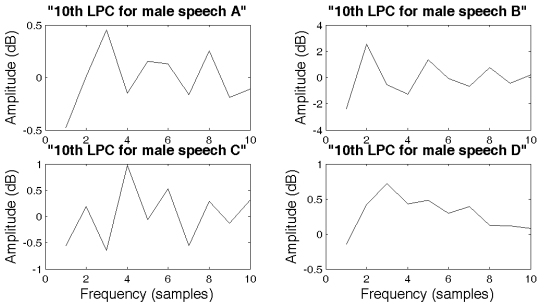

그림 2. 남성의 음성신호에 대한 LPC 계수
Fig. 2 LPC cepstrum coefficients for male speech

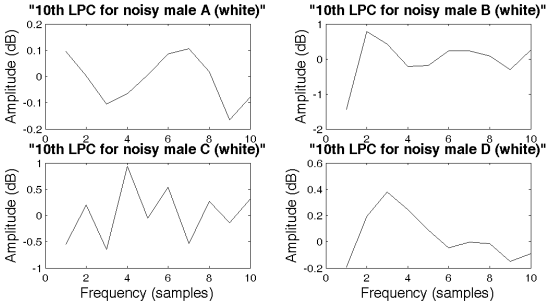

그림 3. 백색잡음이 중첩된 남성에 대한 LPC 계수
Fig. 3 LPC cepstrum coefficients for noisy male speech (addition of white noise)

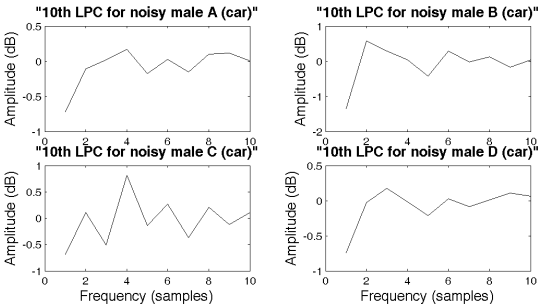

그림 4. 자동차잡음이 중첩된 남성에 대한 LPC 계수
Fig. 4 LPC cepstrum coefficients for noisy male speech (addition of car noise)

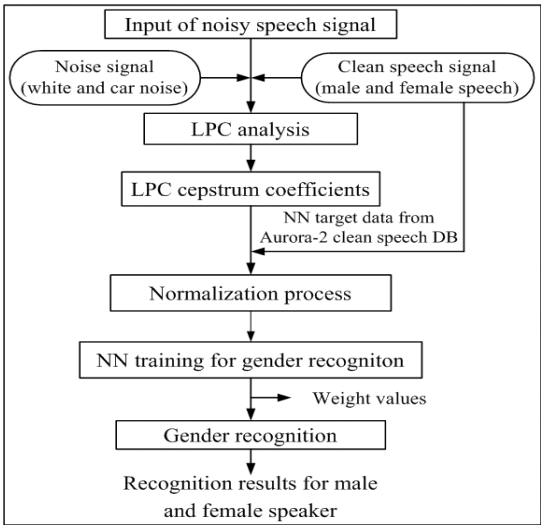

그림 5. 화자종속 인식 알고리즘
Fig. 5 Speaker-dependent recognition algorithm

Ⅳ. 화자인식의 실험결과 및 고찰

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스(Database, DB)[10] 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 (SNR_{in}(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, 12차, 14차의 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 2, 표 3, 표 4에 나타낸다.

표 2. 10차의 LPC 켑스트럼에 대한 성별 인식
Table. 2 Gender recognition for 10 LPC cepstrums

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 3. 12차의 LPC 켑스트럼에 대한 성별 인식
Table. 3 Gender recognition for 12 LPC cepstrums

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%

표 4. 14차의 LPC 켑스트럼에 대한 성별 인식
Table. 4 Gender recognition for 14 LPC cepstrums

Speaker	Gender recognition rates[%]			
	14-20-2 network		14-30-2 network	
	White	Car	White	Car
Male	95.58%	71.68%	91.15%	68.14%
Female	97.35%	97.35%	96.46%	97.35%
Average	96.47%	84.52%	93.81%	82.75%

표에서 알 수 있듯이, 표 2의 10차의 LPC 켑스트럼 계수에 대해서는 10-20-2 네트워크의 인식률이 양호하였으며, 표 3의 12차의 LPC 켑스트럼 계수에 대해서는 12-30-2 네트워크의 인식률이, 표 4의 14차의 LPC 켑스트럼 계수에 대해서는 14-20-2 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 개

선하기 위한 신경회로망의 학습 조건 및 다양한 학습 데이터를 사용할 필요가 있다고 본다. 표로부터 백색잡음에 대해서는 최대 96% 이상의 인식률(14-20-2 네트워크), 자동차잡음에 대해서는 최대 88% 이상의 인식률(10-20-2 네트워크)을 나타내고 있다. 표 5는 본 논문에서 제안한 네트워크의 방법과 기존 방법에 의한 백색잡음 및 자동차잡음 하에서의 화자종속 남녀성별인식에 대한 평균 인식률을 나타낸다.

표 5. 제안한 방법과 기존 방법들에 대한 성별인식률의 비교
Table. 5 Comparisons of the proposed method and conventional method for gender recognition rates

Algorithm	Total gender recognition rates[%]	
	White noise	Car noise
10-20-2	94.69%	88.50%
12-30-2	91.60%	82.30%
RASTA-PLP	95.55%	95.25%
Proposed	96.47%	84.52%

그림에서 "10-20-2"는 표 2의 네트워크에 대한 평균 인식률을, "12-30-2"는 표 3의 네트워크에 대한 평균 인식률을, "RASTA-PLP"는 참고문헌[5]에 대한 평균 성별분류률을, "Proposed"는 본 논문에서 제안한 방식 중에서 가장 양호한 표 4의 10차의 LPC 켑스트럼 계수에 대한 14-20-2 네트워크에 대한 평균 인식률을 각각 나타낸다. 참고문헌[5]의 RASTA-PLP(Relative Spectral Perceptual Linear Predictive) 방법은 피치 성분을 가진 RASTA-PLP 계수를 사용한 성별 분류 시스템이다. 이 시스템에서는 백색잡음 및 자동차잡음에 대하여 0dB, 10dB, 20dB 정도의 잡음만을 가지고 실험을 하고 있기 때문에, 본 실험에서 사용하는 SNR_{in}=-9dB～-15dB(백색잡음) 및 SNR_{in}=-8dB～-16dB(자동차잡음)와 직접 비교하기가 수월하지 않다. 그러므로 본 실험에서 수행한 비슷한 조건으로 맞추기 위하여 SNR_{in}=0dB의 결과와 비교한다. 즉, 본 논문의 SNR_{in}이 최소 -16dB인 반면에 "RASTA-PLP" 방법의 SNR_{in}은 0dB임을 알 수 있다. 본 논문의 알고리즘이 "RASTA-PLP" 방법보다 상당히 낮은 SNR_{in} 환경 하에서도 백색잡음에 대해서는 인식률이 개선되었지만, 자동차잡음에 대해서는 인식률이 떨어지는 것을 알 수 있다. 그러나 낮은 SNR_{in}의 잡음환경 하에서 비교한 것을 고려한다면 어느 정도 개선되었음을 판단할 수 있다. 향후의 연구에서는 이러한 유색잡음에서도 인식률이 개선되도록 다양한 음성, 잡음 데이터에 의한 신경회로망의 개선 노력이 필요하다고 본다. 결론적으로 본 논문의 알고리즘을 이용한 방법이 효과가 있다는 것을 양호한 인식률을 통하여 입증되었다고 판단된다.

V. 결 론

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 실험에서는 3 종류의 LPC 켑스트럼 계수에 대하여 총 6 종류의 신경회로망의 네트워크를 사용하여 제안한 인식알고리즘의 성능평가를 하였다. 백색잡음의 14-20-2 네트워크에 대해서는 최대 96% 이상의 인식률을 구하였으며, 자동차잡음의 10-20-2 네트워크에 대해서는 최대 88% 이상의 인식률을 구하였다. 본 논문에서는 특히 남녀화자의 음성신호에 백색잡음 및 자동차잡음이 중첩된 경우에도 음성인식 실험을 하여, 기존의 방법과 비교하여 본 알고리즘이 효과적인 것을 실험적으로 나타낸다. 또한 SNR_{in}이 낮은 환경 하에서도 특히 백색잡음에 대하여 기존의 방법보다 개선된 결과를 구하였다. 향후 연구 과제로는 좀 더 많은 데이터 셋 및 비정상적인 배경잡음에 대하여 독립적인 화자인식이 가능한 알고리즘을 연구할 예정이다.

참고문헌

[1] A. A. M. Abushariah, T. S. Gunawan, O. O. Khalifa and M. A. M. Abushariah, "English digits speech recognition system based on Hidden Markov Models", 2010 International Conference on Computer and Communication Engineering, pp. 1-5, May 2010.

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

[3] 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.

[4] L. Yang, L. Jing, Y. Yuxiang and W. Jian, "Improvement algorithm of DTW on isolated-word recognition", 2011 IEEE International Conference on Computer Science and Automation Engineering, Vol. 3, pp. 319-322, 2011.

[5] Y. M. Zeng, Z. Y. Wu, T. Falk and W. Y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech", 2006 International Conference on Machine Learning and Cybernetics, pp. 3376-3379, August 2006.

[6] C. C. Chen, P. T. Lu, M. L. Hsia, J. Y. Ke and O.T.-C. Chen, "Gender-to-Age hierarchical recognition for speech", 2011 IEEE 54th International Midwest Symposium on Circuits and Systems, pp. 1-4, 2011.

[7] M. Kos, M.; D. Vlaj and Z. Kacic, "Speaker's gender classification and segmentation using spectral and cepstral feature averaging", 2011 18th International Conference on Systems, Signals and Image Processing, pp. 1-4, 2011.

[8] 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.

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

[10] 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.

저자소개

최재승(Jae-Seung Choi)

1989년 조선대학교 전자공학과 공학사
1995년 일본 오사카시립대학교 전자정보공학부 공학석사
1999년 일본 오사카시립대학교 전자정보공학부 공학박사
2000년～2001년 일본 마쯔시타(현. 파나소닉 주식회사) 전기산업주식회사 AVC사 전임연구원
2002년～2007 경북대 디지털기술연구소 책임연구원
2007년～현재 신라대학교 전자공학과 교수
※관심분야 : 음성신호처리, 적응신호처리, 잡음제거, 신경회로망, 디지털통신, 멀티미디어 등