다층 퍼셉트론 신경회로망을 이용한 후두 질환 음성 식별

강 현 민, 김 유 신, 김 형 순
부산대학교 전자공학과

Detection of Laryngeal Pathology in Speech Using Multilayer Perceptron Neural Networks

Hyun Min Kang, Yoo Shin Kim, Hyung Soon Kim
Department of Electronics Engineering, Pusan National University
E-mail : kanghm@pusan.ac.kr

Abstract

Neural networks have been known to have great discriminative power in pattern classification problems. In this paper, the multilayer perceptron neural networks are employed to automatically detect laryngeal pathology in speech. Also new feature parameters are introduced which can reflect the periodicity of speech and its perturbation. These parameters and cepstral coefficients are used as input of the multilayer perceptron neural networks. According to the experiment using Korean disordered speech database, incorporation of new parameters with cepstral coefficients outperforms the case with only cepstral coefficients.

Ⅰ. 서 론

일반적으로 후두 질환이 있는 사람의 음성은 거칠고 쉰 목소리가 많다. 특히, 이러한 현상은 모음 부분에서 두드러지는데, 이는 후두에 있는 발성 기관인 성대에 문제가 생기기 때문이다. 따라서, 환자의 음성 청취는 후두 질환을 검진하는데 중요한 도구가 되며, 후두 질환의 유무 판단을 자동적으로 해 보고자 하는 것이 본 논문의 목적이다.

후두 질환을 감별해 내기 위해서는 후두의 상태를 잘 반영할 수 있는 음성 변수들을 찾아내는 일이 아주 중요한데, 이와 관련한 연구로는 피치의 동요 요인(pitch perturbation factor)을 사용한다거나[1][2], 후두 질환 음성에서 잡음 성분을 이용하는 방법 등이 있었다[3][4]. 이외에도 음성 강도의 변화를 살펴보거나[5][6], 최근에는 켑스트럼 계수를 활용하는 방법이 활발히 연구되고 있다[7][8].

그렇지만, 일반적인 패턴 인식 문제와 마찬가지로 한 종류의 파라미터로 성취할 수 있는 인식률은 한계가 있게 마련이고, 서로 독립적이면서도 많은 정보를 줄 수 있는 파라미터의 조합을 연구하는 것이 필요하다.

본 논문에서는 음성인식에 널리 사용되는 켑스트럼 계수를 하나의 파라미터 군으로 하고 주기성 정도와 그 주기성의 동요 정도를 나타내는 또 다른 특징 파라미터 군을 함께 사용한다. 주기성에 관한 파라미터는 자기 상관 함수를 사용하여 얻어낸다. 그리고, 이 파라미터들을 입력으로 하여 다층 퍼셉트론 신경회로망이 후두 질환 음성을 분류해 내는 역할을 한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 실험에 사용한 음성 식별 파라미터를 소개하고, 신경회로망을 포함한 전체 시스템의 구성을 살펴본다. 3장에서는 실험 방법과 결과를 보여주고, 마지막으로 4장에서는 향후 연구 방향 제시와 함께 결론을 맺는다.

Ⅱ. 후두 질환 음성 식별 시스템

음성의 식별은 프레임 단위로 이루어지는데, 식별에

사용한 파라미터는 크게 두 가지로 나누어진다. 첫 번째 파라미터 부류는 켑스트럼 계수로써 이 변수들은 각 프레임 단위로 음성의 특성을 나타낸다. 그 중에서 LPCC(Linear Predictive Cepstral Coefficient)는 사람의 발성 기관의 특성을 고려한 파라미터이고 MFCC(Mel-Frequency Cepstral Coefficient)는 사람의 청각 기관의 특성을 고려한 파라미터이다. 이들 두 파라미터는 모두 음성 인식 분야의 대표적인 특징 추출법으로 인정되고 있다[9].

두 번째 파라미터 부류는 음성의 주기성이 한 프레임 내에서 얼마나 강한지, 또는 음성의 주기(피치)가 일정 기간 동안 얼마나 일정하게 유지되는지를 알아보도록 정하였다.

n 번째 프레임 내에서의 자기 상관 함수를 $R_n(k)$로 나타낼 때, 주기성 정도, $V(n)$는 식(1)과 같이 구하였다.

$$V(n) = \max_{k_{\min} \le k \le k_{\max}} R_n(k)/R_n(0) \qquad (1)$$

여기서, $k_{\min}$과 $k_{\max}$는 각각 사람의 피치 주기가 존재할 수 있는 최저 위치와 최고 위치를 나타낸다.

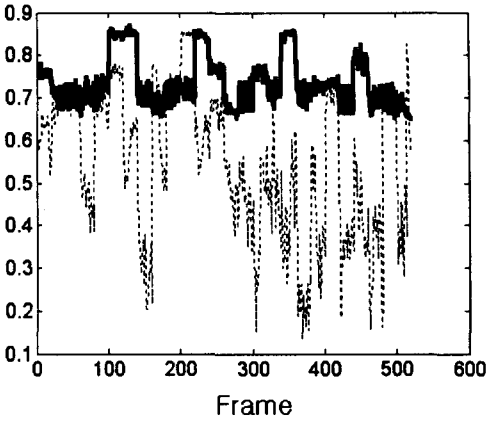
그리고, 현재 프레임과 중복되지 않고 가장 가까이 인접한 좌우 각각 두 프레임과 현재 프레임을 포함하여 다섯 프레임에 있어서 $V(n)$의 분산을 주기성의 동요 정도에 대한 파라미터 $VAR_V(n)$로 정하였다.

$$VAR\_V(n) = \frac{1}{5}\sum_{k=-2}^{2}\left[V(n+k) - \frac{1}{5}\sum_{l=-2}^{2}V(n+l)\right]^2 \qquad (2)$$

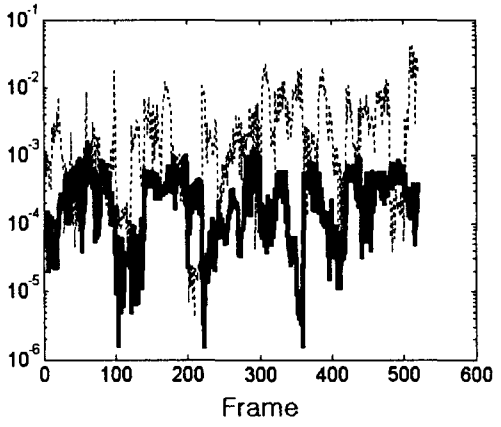
그림 1에서는 주기성과 관련된 파라미터들이 적절한 파라미터가 될 수 있음을 보여주고 있다. 그림은 학습과 인식에 사용되는 여러 사람들의 프레임들에서 나온 결과를 보기 좋게 연결하여 놓은 것이다.

그림 2는 악성 종양 환자와 정상인의 음성 파형을 비교해 놓은 그림인데, 후두 질환 환자의 음성은 주기적인 특성이 정상인에 비해서 덜 분명함을 알 수 있다. 그림 3은 자기 상관 함수가 음성의 주기성을 판단하는 좋은 도구가 됨을 보여주는 그림이다. 주기적인 신호일 경우 자기 상관 함수도 주기적인 경향을 띠게 되는데, 그림에서 확인이 되듯이 정상인의 경우가 후두 질환이 있는 사람의 경우보다 훨씬 더 주기적인 경향이 강함을 알 수가 있다.

인식기는 다층 퍼셉트론 신경회로망을 사용하였고, 학습 규칙은 일반적인 오차 역전파 알고리즘을 이용하였다[10]. 그림 4는 전체적인 시스템의 구성을 간략하게 보여준다.



(a)



(b)

그림 1. 주기성 파라미터들의 수치비교. 실선은 정상인의 경우이고, 점선은 악성 종양 환자의 경우임. (a) $V(n)$ 값들 (b) $VAR\_V(n)$ 값들 (로그 스케일)

III. 인식실험 및 결과

본 실험을 위해 한국 장애 음성 데이터베이스 v1.0을 사용하였다[11]. 이 데이터베이스에서 음성에 대한 녹음과 환자에 대한 임상 정보의 수집은 부산대학교 의과대학 이비인후과에서 행해졌으며, 수집된 정보의 분류 및 정리는 창원대학교 제어계측공학과 신호 및 시스템 연구실에서 수행되었다. 음성 녹음은 방음실에서 이루어졌으며, 녹음 대상자를 편안한 자세로 앉게 한 후 마이크 앞에서 15cm 가량 거리를 두고 평상시와 같은 음높이로 약 3초간 /아/, /이/, /우/, /에/, /오/를 발성하도록 하였다. 샘플링 주파수는 16kHz이며,

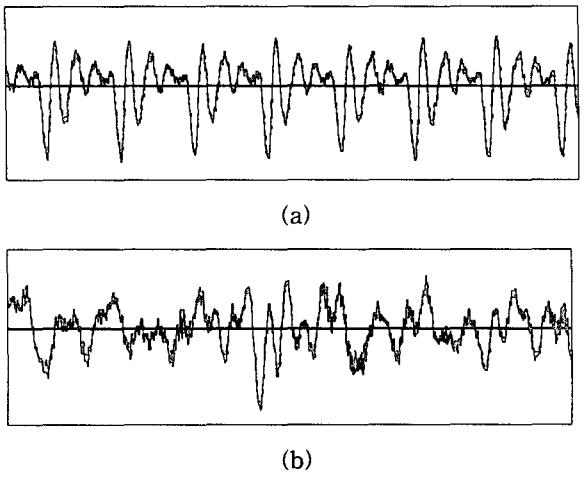


그림 2. 정상인과 악성 종양 환자의 음성 파형의 예 (모음 /아/) (a) 정상인의 음성 (b) 악성 종양 환자의 음성

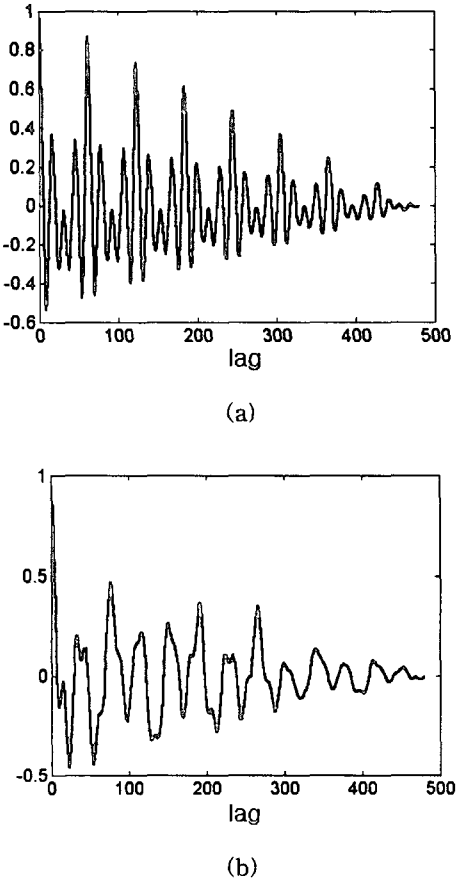


그림 3. 정상인과 악성 종양 환자의 정규화된 자기 상관 함수 $R_n(k)/R_n(0)$의 예 (a) 정상인의 정규화된 자기 상관 함수 (b) 악성 종양 환자의 정규화된 자기 상관 함수

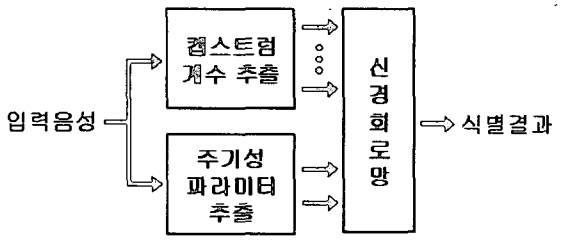


그림 4. 후두 질환 음성 식별 시스템의 구성

양자화 비트 수는 16비트를 사용하였다. 전체 음성 데이터는 악성 종양 환자의 음성 31개, 양성 종양 환자의 음성 28개, 그리고 정상인의 음성 41개로 구성되어 있으나, 실험 대상으로는 동일하게 각각 26명의 음성 파일을 실험 대상으로 삼았다.

전체 음성 파일 중에서 실험에 사용한 부분은 /아/ 음성 부분으로, 파형 편집기로 직접 손으로 따로 떼어내어 사용하였다. 서론에서도 언급하였듯이 후두의 손상으로 음성에 생기게 되는 변화는 모음 부분에서 뚜렷이 나타나므로, 대표적인 모음으로 /아/ 음성 부분을 채택하였다. 그러나, 여러 모음을 고려하거나 음성이 변하는 부분 또한 연구하여야 할 향후 과제가 될 수 있을 것이다. 안정적인 모음 발성 부분을 얻어내기 위해서 /아/ 음성의 중앙 지점 20 프레임을 실험에 사용하였다. 즉, 한 사람당 20 프레임의 데이터를 얻고, 전체적으로는 악성 520 프레임, 양성 520 프레임, 그리고, 정상 520 프레임의 데이터를 얻게 된다. 프레임은 30ms의 길이를 가지고, 각 프레임 간의 중복 구간의 길이는 20ms로 하였다. LPCC나 MFCC를 얻기 위해서 프레임에 Hamming 윈도우를 사용하였으며, pre-emphasis 수치는 0.97로 하고, 켑스트럼 계수의 차수는 둘 다 12차로 하였다.

신경회로망은 은닉층을 하나 가지고 있는 다층 퍼셉트론 네트워크를 사용하였고, 은닉층 노드 수는 10개, 출력층 노드 수는 1개, 학습 계수는 0.01, 그리고, 학습 횟수는 전체 학습 데이터를 한 번 거치는 것을 학습 횟수 한 번으로 정하여 3000번으로 고정시켰다. 입력층의 노드는 켑스트럼 파라미터만 사용할 경우에는 12개, 주기성 파라미터도 함께 사용할 경우에는 14개를 사용하였으며, 각 노드의 입력 값들은 평균이 0, 분산이 1이 되도록 정규화 시켰다. 전체 26명 중에 17명을 학습 데이터로 하고, 나머지 9명을 시험 데이터로 사용하였다.

그런데, 비록 한 사람에게서 20개의 프레임이 얻어지더라도 그 20개 프레임은 거의 비슷한 특성을 가지

는 데이터라서 실제로는 데이터가 부족한 편이었다. 그래서, 어떤 사람들을 학습 데이터와 시험 데이터로 삼느냐에 따라 인식 결과의 변동이 심하였다. 이에 대한 해결 방안으로 신경회로망의 학습과 시험을 100번 반복하여 그 결과의 평균을 취하되, 반복할 때마다 학습 데이터로 선택되는 사람과 시험 데이터로 선택되는 사람을 난수적으로 선택하도록 하였다.

표 1은 실험 결과의 인식률을 나타낸다. 실험은 악성과 정상을 구분하는 것, 악성과 양성을 구분하는 것, 그리고, 양성과 정상을 구분하는 것을 각각 실행해 보았다. 예상대로 악성과 정상을 구분하는 인식률이 가장 높게 나왔으며, 주기성을 나타내는 파라미터가 함께 사용될 때, 더 높은 인식률을 가져옴을 확인할 수가 있었다.

표 1. 인식률 실험 결과

	악성/정상	악성/양성	양성/정상
LPCC	79.03%	74.80%	63.65%
LPCC+주기성	88.71%	77.73%	66.54%
MFCC	81.35%	73.31%	68.30%
MFCC+주기성	88.40%	74.10%	69.17%

IV. 결 론

본 논문에서는 사람의 음성으로부터 특징 파라미터들을 추출한 다음 신경회로망을 이용하여 자동적으로 후두 질환 여부를 감별해 내는 방법을 검토하였다. 그리고, 특징 파라미터로 주기성의 정도와 주기성의 요동 정도를 표현해 주는 파라미터를 도입하였다. 음성인식에 널리 사용되는 LPCC와 MFCC가 후두 질환의 특성을 나타내는 파라미터가 될 수 있음을 확인하였고, 두 파라미터 모두 주기성 파라미터와 함께 사용하였을 때, 더 좋은 결과를 얻을 수가 있었다. 아직은 성능이 미흡한 편이므로, 앞으로 개서된 연구 결과를 얻기 위해서는 후두 질환 특성을 보다 잘 표현해 줄 수 있는 새로운 특징 파라미터의 도입이 필요하다고 판단된다.

본 논문은 보건복지부 협동기초연구지원 연구개발 사업 연구결과의 일부입니다.

참고문헌

[1] P. Lieberman, "Perturbations in vocal pitch," *J. Acoust. Soc.* Am 33: pp. 597-603, 1961.

[2] S. Iwata, "Periodicities of pitch perturbation in normal and pathologic larynxes," *Laryngoscope* 82: pp. 87-96, 1972.

[3] E. Yumoto, W. J. Gould, T. Baer, "Harmonic-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc.* Am 71: pp. 1544-1550, 1982.

[4] H. Kasuya, S. Ogawa, K. Mashima, S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *J. Acoust. Soc.* Am 80(5), Nov., 1986.

[5] Y. Koike. H. Takhashi, T. C. Calcatera, "Acoustic measurements for detecting laryngeal pathology," *Acta Otolaryngol*, 85: pp. 105-107, 1977.

[6] Y. Horri, "Jitter and Shimmer in sustained vocal fry phonation," *Folia Phoniatrica*, vol. 37, pp. 81-86, 1985.

[7] C. E. Martinez, H. L. Rufiner, "Acoustic analysis of speech for detection of laryngeal pathologies," *IEEE Int. Conf. EMBS*. pp. 2369-2372, 2000.

[8] 김용주, "음성분석과 인식기법을 이용한 후두질환 식별 파라미터 개발," *부산대학교 석사논문*, 2002년 2월.

[9] L. Rabiner, B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[10] J. M. Zurada, *Introduction to Artificial Neural Systems*, Web Publishing Company, 1992.

[11] Korean Disordered Speech Database, Version 1.0, 창원대학교, 1999년.