주성분 분석을 이용한 최적 흉부음 데이터 검출

*임 선 희, **박 기 영, ***최규훈, ****박 강 서, *김 종 교
*전북대학교 전자정보공학부, **전주공업대학 정보통신과,
***전주공업대학 전자정보과, ****전주예수병원 소아과
전화 : 063-272-1177

Optimal Thoracic Sound Data Extraction Using Principal Component Analysis

*Sun-Hee Lim, **Ki-Young Park, ***Kyu-Hoon Choi,
****Kang-Seo Park, *Chong-Kyo Kim

*Division of Electronics & Information Engineering, Chonbuk National University
**Dept. of Information & Communication Engineering, Jeonju Technical College
***Dept. of Electronics & Information Engineering, Jeonju Technical College
****Dept. Pediatrics, Jeonju Jesus Hospital
E-mail : sponzi@hanmail.net

Abstract

Thoracic sound has been widely known as a good method to examine thoracic disease. But, it's difficult to diagnose with correct data according to patient's thoracic position from same patient who has thoracic disease. Therefore, it is necessary to normalize the data for lung sound objectively. In this paper, we'd like to detect a useful data for medical examination by applying PCA(Principal Component Analysis) to thoracic sound data and then present a objective data about lung and heart sound for thoracic disease.

1. 서 론

지금까지 사람의 흉부에서 들리는 소리(생체음)는 청진기를 이용한 청각만을 이용하여 들을 수 있었다. 그러나 생체음은 같은 환자임에도 불구하고 상황에 따라 다르게 소리나고 듣는 사람에 따라 다르므로 오랜 경험과 노하우를 가진 의사들만이 비교적 정확한 진단을 내릴 수 있는 반면 경험이 적은 의사들은 진단에 애로가 많았다.

본 논문에서는 주성분 분석을 이용하여 측정한 흉부 질환 생체음 즉, 흉부음에서 질환 판별에 유용한 정보를 검출한 것이다.

흉부질환 생체음 데이터에 주성분 분석(PCA:Principal Component Analysis)을 적용 하여 주성분 데이터를 검출하고, 이때 주성분으로 검출된 데이터는 진단에 유용한 정보의 손실없는 고유 흉부음 즉, 최적의 흉부 질환 생체음을 검출할 수 있다.

검출된 최적의 흉부 질환 생체음 데이터와 원(original)흉부 생체음 데이터를 비교 분석하고 전문의를 통하여 최적의 흉부 생체음에 대한 객관적인 평가를 제시한다.

2. 주성분 분석을 이용한 흉부음 분석

2.1 흉부음과 흉부 질환

흉부음(thoracic sound)은 정상폐에서 폐포에 공기가 이동하면서 나는 low pitch sound로서 주로 숨을 들이마시는 경우에 들리게 되는 폐포 호흡음과, high pitch sound의 기관지 호흡음으로 분류할 수 있다[1]. 이때 흉부 질환을 가진 환자는 이 호흡음에 잡음이 더해진 경우가 된다.

흉부 질환을 갖는 생체음에는 천명음(wheeze), 수포음(rale), 통음(rhonchi)의 세 가지를 들 수 있다. 이중 천명음은 호흡시 발생하는 음악적인 휘파람 소리로, 이 천명음이 발생되는 것은 기도의 어느 부위에 기류

가 막혀서 나는 소리이다. 천명음은 호흡곤란과 연관되어 나타나며, 대부분 천식 환자의 상태가 악화될 때 천명음이 나타나며, 천식(asthma)은 재발성 천명음의 가장 흔한 원인이다. 수포음은 비교적 작은 기관지 또는 폐포에서 발생되는 바삭바삭하는 음(crackling sound)을 나타낸다. 이런 음이 발생하는 경우는 폐렴(pneumonia)인 경우이다. 통음은 연속적으로 들리는 악음으로서 코고는 소리(snoring sound) 또는 덜커덕거리는 소리(rattling sound)로 표현되며, 이는 기관지염(bronchitis)의 존재를 시사한다[2].

2.2 주성분 분석

주성분 분석(PCA)은 서로 연관된 변수들의 집합을 간결하게 서술하기 위해서 수행되어진다. 이 기술은 서로 연관된 원시변수들을 상호 연관성이 없는 새로운 변수들로 변환하는 방법으로 요약할 수 있다. 그러한 새로운 변수들을 주성분이라 하고, 각 주성분이 전달하는 정보량은 주성분의 분산으로 측정하며, 분산을 최대화하는 주성분을 제1주성분(first principal component)이라고 한다.

어떤 연구자는 문제에 대한 차원을 감소시키는 것, 즉 많은 정보를 소실하지 않으면서 변수들의 수를 감소시키기를 원할 것이다. 이것은 단지 몇 개의 중요한 주성분만을 사용함으로써 가능하다. 주성분 분석은 변수의 수를 감소시킬 뿐만 아니라 주성분들이 상호 연관성이 없으므로 매우 유용하게 사용되어진다. 그러므로 주성분 분석을 이용할 경우 복잡한 상호연관성을 가지는 많은 수의 원시변수들을 분석하는 대신에 상호 연관성이 없는 적은 수의 주성분을 분석함으로써 문제를 해결할 수 있다.

고유 벡터(eigenvector)를 생성하는 과정은 다음과 같다.

1) 공분산 행렬을 생성한다.

학습데이터의 $i(=1..n)$번째 데이터를 나타내는 열벡터를 x_i 라 하면 행렬 $X(=[x_1|x_2|..x_n])$을 만들 수 있다. 행렬 X에 대한 공분산 행렬은 행렬 연산 $\Sigma_X = X^* X^T$으로 구할 수 있으며, Σ_X 학습 데이터의 i 와 j 의 공분산을 표현한다.

2) 공분산 행렬의 고유값과 고유벡터를 구한다.

열벡터 x_i를 주성분 공간에서 표현하는 열벡터를 y_i라 하면 행렬 $Y(=[y_1|y_2|..|y_n])$는 주성분 공간에서 학습데이터를 포함하는 행렬이다. 따라서 행렬 Y가 포함하는 데이터는 서로 연관되어 있지 않으므로, 그것에 대한 공분산 행렬 $\Sigma_Y(=Y^* Y^T)$는 대각행렬이 되어야 한다. 주성분은 선형적으로 계산되어질 수 있다. 그것의 열벡터들이 정규직교 $(P^T * P = I)$하는 변환행렬을 P라 하면, 행렬 X와 Y에 대해 식(1), (2)가 성립한다.

$$Y = P^T * X \text{ and } X = P^* Y \quad (1)$$

$$\Sigma_Y = Y^* Y^T = P^T * \Sigma_X^* P \quad (2)$$

공분산 행렬 X에 대한 고유벡터를 포함하는 행렬 P와 고유값을 포함하는 대각행렬에 대해 다음의 식(3), (4)를 만족한다.

$$\Sigma_X^* P = \Lambda^* P \quad (3)$$

$$\Sigma_Y = P^T * \Lambda^* P = \Lambda^* P^T * P = \Lambda \quad (4)$$

즉, 고유벡터를 구하는 것은 위의 식을 만족하는 고유값 행렬 Λ와 고유값에 대한 고유벡터 P를 구하는 것이다[3][4][5].

2.3 주성분 분석에 의한 흉부음 분석

각 N개의 데이터 열로써 구성된 흉부음 데이터는 N차원의 벡터 공간이라고 생각할 수 있다. 이 데이터 열에 대한 주성분 분석의 주요 개념은 전체 데이터 공간 안에서 흉부음의 분포를 가장 잘 계산하는 소수의 벡터를 찾아내는 것이다. 이들 벡터들을 흉부음 공간 좌표축으로 정의하고 이것을 본 논문에서는 흉부음 공간이라고 한다. 각 벡터들의 길이는 각 흉부음 데이터의 구성 샘플 개수인 N이다. 이 벡터들은 원 흉부음 데이터에 대응하는 공분산 매트릭스의 고유벡터들인데 원래 흉부음 데이터와 비슷한 형상을 가지므로 본 논문에서는 이 벡터들을 고유 흉부음이라고 부르기로 한다.

본 논문에서는 흉부 질환에 따른 고유 흉부음을 얻었는데 그 과정은 먼저 흉부음 데이터 세트로부터 공분산 행렬를 구한다. 그리고 고유값과 고유벡터들을 구하고 고유값의 순서대로 순차적으로 정렬한다. 이 고유벡터들은 흉부음 데이터 간의 변동을 특징짓는 특징 세트로 간주된다[6].

2.3.1 고유 흉부음 추출

M개의 흉부음 데이터의 세트를 각각 $X_1, X_2, ..., X_M$라고 하자. 세트의 평균 흉부음을 m_x라고 하고 아래 식(5)와 같이 정의한다.

$$m_X = \frac{1}{M}\sum_{n=1}^{M} X_n \quad (5)$$

하나의 흉부음 데이터는 N개의 샘플 데이터들로 구성되어 있다면 1개의 심음 데이터 X는 다음 식(6)과 같이 표현된다.

$$X = \sum_{n=1}^{N} x(n) \quad (6)$$

본 논문에서 한 흉부음의 데이터 샘플수 N은 34000이고 한 병명 당 훈련 세트는 10이다.

다음 단계는 흉부음 벡터 X 들의 집합에 대한 행렬을 구한다. 공분산 행렬은 다음 식(7)과 같다.

$$C_X = \frac{1}{M}\sum_{n=1}^{M}(X_n X_n^T - m_n m_n^T) \quad (7)$$

이 공분산 행렬부터 식(3), (4)을 만족시키는 고유치를 구한 후 고유벡터를 구하면 우리가 원하는 고유 흉부음을 얻을 수 있다. 그림 1은 고유 흉부음을 추출하는 블록도이다.

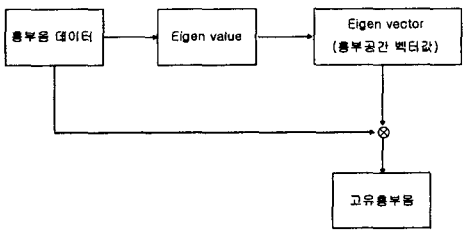

그림 1. 고유 흉부음 추출 블록도

3. 실험 방법 및 결과

본 논문의 실험을 위해서 사용한 생체음 데이터는 전주예수병원에서 직접 환자로부터 얻은 데이터로서 조용한 환경에서 녹음된 데이터이다. 각 질환 당 10명의 환자로부터 얻은 흉부 질환 생체음 데이터베이스를 구축하여 실험을 하였다.

그림 2, 3, 4는 각각 천식, 폐렴, 기관지염 환자의 흉부 생체음 세트 10개의 정규화된 데이터를 나타내었다. 각 그림에서 가로축은 샘플 수이며 세로축은 각 샘플의 크기이다. 이하 그림들도 가로축과 세로축의 정의는 동일하다.

그림 5, 6, 7은 각각 천식, 폐렴, 기관지염 질환에 대한 고유 흉부음을 나타냈다. 그림 5에서 (1)번 그림이 각 흉부 질환 데이터들의 공통 성분을 가장 많이 포함하고 있는 그림이며 나머지 부분에서 공통 성분을 가장 많이 포함하고 있는 그림은 (2)번 그림이다. 같은 원리로 (10)번 그림이 공통 성분을 가장 적게 포함하고 있다. 즉, (1)번 그림이 흉부 질환 데이터 세트에서 가장 중요한 고유 흉부음이 된다. 그림 6, 7도 마찬가지이며, 홀수 번의 그림만 나타냈다.

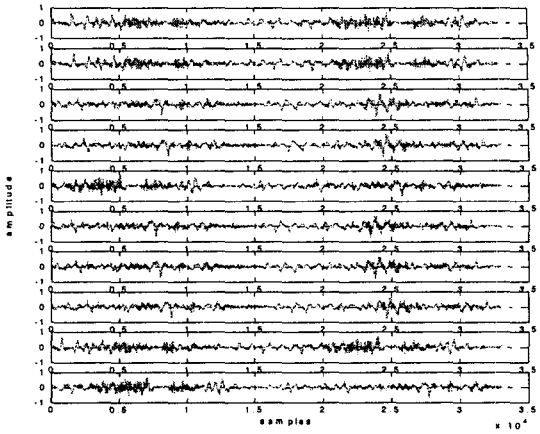

그림 2. 천식 환자의 흉부음 세트

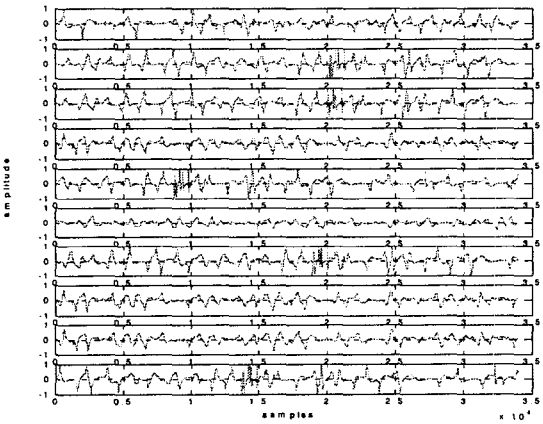

그림 3. 폐렴 환자의 흉부음 세트

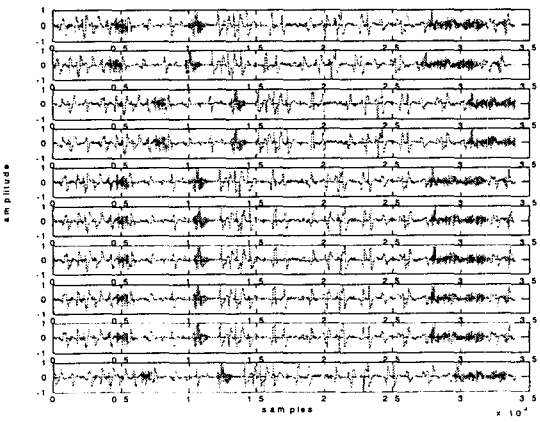

그림 4. 기관지염 환자의 흉부음 세트

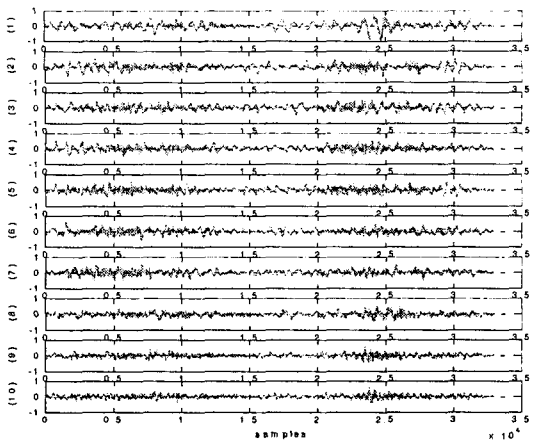

그림 5. 천식 환자의 고유 흉부음

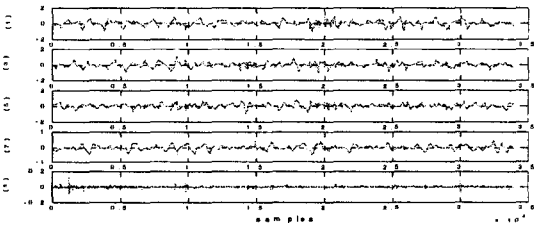

그림 6. 폐렴 환자의 고유 흉부음

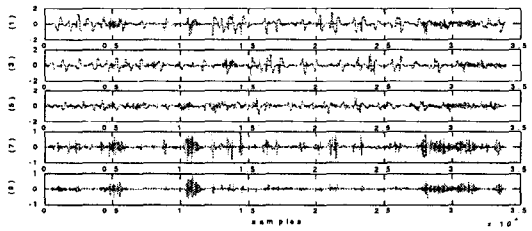

그림 7. 기관지염 환자의 고유 흉부음

표 1. 흉부음 공간의 첫 번째 주성분 벡터값

	천식	폐렴	기관지염
흉부음공간 벡터값	-0.0787	-0.0654	0.0937
	-0.0355	0.1639	0.0570
	0.5239	0.1310	0.0298
	-0.3830	-0.5752	0.0240
	0.0023	-0.0257	-0.4490
	-0.4130	-0.2570	-0.1525
	0.4845	-0.0469	-0.1973
	0.3854	-0.5677	-0.5996
	-0.0582	-0.5431	-0.6011
	0.1194	0.0020	0.0449

표 1은 각각의 흉부 질환 데이터 세트에서 가장 중요한 그림 5, 6, 7에서 (1)번째 고유 흉부음을 결정짓는 벡터 값을 나타냈다.

4. 결 론

본 논문에서는 흉부 질환 생체음의 객관적 자료화를 위해 주성분 분석을 이용하여 고유 흉부음을 구하였다. 그 결과 각 질환에 대해 진단에 유용한 생체음 데이터를 검출할 수 있었다.

검출한 고유 흉부음을 진문의사에게 들려주고 평가하였을 경우 만족할 만한 평가를 얻었고, 각 병명을 결정짓는 특징 데이터 검출하는 결과를 얻을 수 있었다.

질환을 갖는 흉부음 특성상 잡음 성분이 강하기 때문에 그림 5, 6, 7에서 (2), (3)번째 고유 흉부음도 병명을 판단하는 데는 적합한 것으로 나타났다.

향후, 각 흉부질환별 상세한 DB 구축 작업이 수반된다면 좀더 최적의 흉부 질환 생체음 추출에 도움이 될 것이다.

본 연구는 산업기술자원부의 공통핵심기술개발 사업의 일부 지원으로 이루어졌다.

참고문헌

[1] 신영기, *임상진단학*, 계축문화사, 1993.

[2] Mark H. Beers, Robert Berkow, *The Merck Manual of diagnosis and therapy*. Merk Research Laboratories Division of Merck & Co. INC Whitehouse Statio N. J. 2000.

[3] Sami Romdhani, "Face Recognition Using Prin cipal Components Analysis", Master thesis, http://www.elec.gla.ac.uk/~romdhani.

[4] H. Hotelling, "simplified calculation of principal companents", Psychometricka, vol. 1, pp. 27-35, 1936.

[5] S. J. Lee, S. B. Jung, J. W. Kwon and S. H. Hong, "Face Detection and Recognition Using PCA", *TENCON'99*, vol. 1, pp. 84-87, Cheju, 9. 1999.

[6] 이상민, "재샘플링과 주성분 분석에 의한 시간-주파수 영역에서의 심음 인식", 박사학위논문, 인하대학교, 2000.