후두질환 음성의 자동 식별 성능 비교*

강현민(부산대), 김수미(부산대), 김유신(부산대), 김형순(부산대)
조철우(창원대), 양병곤(동의대), 왕수건(부산대)

<차 례>

<Abstract>

Performance Comparison of Automatic Detection of Laryngeal Diseases by Voice

Hyun Min Kang, Soo Mi Kim, Yoo Shin Kim, Hyung Soon Kim, Cheol-Woo Jo, Byunggon Yang, Soo-Geun Wang

Laryngeal diseases cause significant changes in the quality of speech production. Automatic detection of laryngeal diseases by voice is attractive because of its nonintrusive nature. In this paper, we apply speech recognition techniques to detection of laryngeal cancer, and investigate which feature parameters and classification methods are appropriate for this purpose. Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) are examined as feature parameters, and parameters reflecting the periodicity of speech and its perturbation are also considered. As for classifier, multilayer perceptron neural networks and Gaussian Mixture Models (GMM) are employed. According to our experiments, higher order LPCC with the periodic information parameters yields the best performance.

* Keywords: laryngeal disease, cepstrum, periodicity, multilayer perceptron, GMM

* 본 논문은 보건복지부 협동기초연구지원 연구개발사업 연구 결과의 일부입니다 (02-PJ1-PG10-31401-005)

1. 서 론

후두질환은 환자의 음성 특성에 큰 변화를 가져오며, 일반인보다 거칠고 쉰 목소리가 나는 증상 등이 그 예이다. 환자의 음성 청취는 후두 질환을 검진하는데 중요한 도구가 되며, 최근 신호처리 기술의 발달과 더불어 음성신호의 자동분석에 의한 후두질환 식별에 많은 관심이 모아지고 있다. 음성분석에 의한 후두질환 식별은 식별 성능 만 어느 정도 이상 보장된다면, 고통 없이 검사를 신속 간편하게 할 수 있을 뿐만 아니라, 인터넷 등을 통한 원격 검진이 가능하다는 장점을 지닌다.

본 논문에서는 음성인식 기술을 기반으로 한 후두질환 음성의 자동 식별 방식들을 검토하고 그 성능을 비교하였다. 후두질환 음성의 식별을 위해서는 먼저 후두의 상태를 잘 반영할 수 있는 음성 파라미터들을 찾아내는 일이 중요하다. 이와 관련된 연구로는 피치의 동요 요인(pitch perturbation factor)을 사용하거나[1][2], 후두 질환 음성에서 잡음 성분을 이용하는 방법 등이 있었다[3][4]. 이외에도 음성 강도의 변화를 이용하기도 하고[5][6], 최근에는 켑스트럼 계수를 활용하는 방법이 연구되고 있다[7][8].

본 논문에서는 후두암 음성 식별을 위한 기본적인 특징 파라미터로 음성인식에 널리 사용되는 켑스트럼 계수인 MFCC (Mel-Frequency Cepstral Coefficients)와 LPCC (Linear Predictive Cepstral Coefficients)를 사용하였다. 이들이 주파수 영역에서 음성 스펙트럼의 포락선을 표현해 주는 파라미터이므로, 음성의 주기적인 특성을 충분히 반영해 주지는 못한다. 따라서, 본 논문에서는 시간 영역에서 음성의 주기성 정도와 그 주기성의 동요 정도를 나타내 주는 특징 파라미터들을 함께 사용하도록 하였다. 인식기로는 다층 퍼셉트론 신경회로망과 Gaussian Mixture Model (GMM) 기반의 인식기를 함께 검토하며, 상기 특징 파라미터들과 인식기들에 의한 정상인과 후두암 환자의 음성을 자동 식별하는 실험을 수행하여 그 성능을 비교하였다.

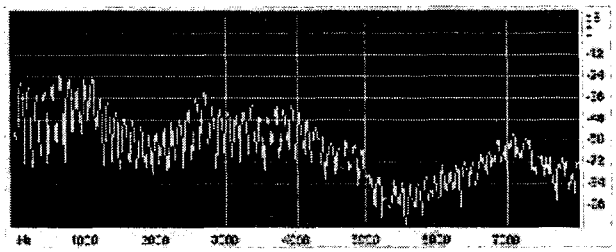
본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 실험에 사용한 음성 식별 파라미터를 소개하고, 3장에서는 신경회로망을 포함한 전체 시스템의 구성을 살펴본다. 4장에서는 음성 데이터와 실험 방법, 그리고 실험 결과를 다루고, 마지막으로 5장에서 결론을 맺는다.

2. 후두질환 식별을 위한 특징 파라미터

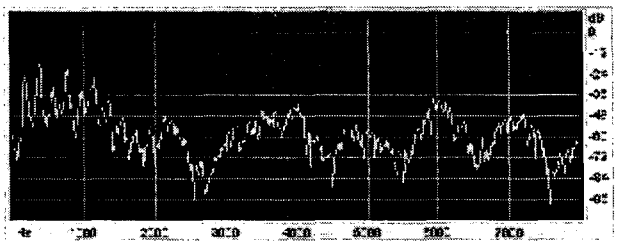
2.1. 켑스트럼 계수

<그림 1>은 정상인과 후두암 환자가 단모음 /아/를 발음했을 때의 음성 스펙트럼을 보여준다. 후두암 환자의 음성은 잡음 성분이 많으므로 스펙트럼의 고주파 성분이 상대적으로 높고, 주기성이 떨어지므로 주파수 영역에서의 주기적인 하모닉 성분들이 잘 드러나지 않는다. 이 그림에서 음성 스펙트럼의 포락선 정보만으로도 정상인과 후두암 환자의 구별이 가능함을 알 수 있다.

켑스트럼 분석은 음성신호의 로그 스펙트럼의 역 Fourier 변환에 기반한 것으로서, 20차 이내의 낮은 차수의 켑스트럼 계수들은 음성 스펙트럼의 포락선 정보를 표현해 준다. 음성인식에 널리 사용되는 켑스트럼 계수로 LPCC(Linear Predictive Cepstral Coefficients)와 MFCC(Mel-Frequency Cepstral Coefficients)가 있다[9]. LPCC는 음성발생 기관의 선형예측 모델에 기반을 둔 것이고, MFCC는 청각 기관의 주파수 특성을 고려한 파라미터이다. 본 논문에서는 정상인과 후두암 환자의 음성의 스펙트럼 포락선 특성을 구별하기 위한 특징 파라미터로 LPCC와 MFCC를 적용하고, 이들의 성능을 비교한다.



(a)



(b)

<그림 1> 정상인과 후두암 환자의 음성 스펙트럼의 예 (모음 /아/)
(a) 정상인의 음성 (b) 후두암 환자의 음성

2.2. 주기성 정보

<그림 2>는 정상인과 후두암 환자의 대표적인 음성 파형 일부분을 서로 비교해 놓은 그림이다. 그림을 보면 정상인의 음성이 후두암 환자의 음성에 비해 주기적인 특성이 훨씬 더 분명함을 알 수가 있다. 따라서 주기성의 정도를 적절히 표현하면 후두암 환자의 음성을 식별할 수 있는 유용한 파라미터가 될 것이다. 본 논문에서는 음성신호의 자기상관함수를 이용하여 시간 영역에서 주기성의 정도를 측정하였다. 주기적인 신호일 경우 자기상관함수도 주기적인 경향을 나타내므로, <그림 3>에서 보는 바와 같이 정상인 음성의 자기상관함수가 후두암 환자의 경우보다 주기적인 경향이 강하게 된다. 본 논문에서는 자기상관함수를 이용한 다음 두 가지 특징 파라미터를 이용하였다.

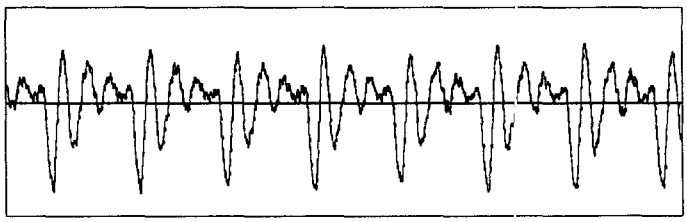
첫 번째 파라미터는 한 프레임 내에서의 음성의 주기성에 관한 것으로, n번째 음성 프레임에서의 자기상관함수를 Rn(k)로 나타낼 때, 주기성 정도인 V(n)는 다음 식과 같이 나타낸다.

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

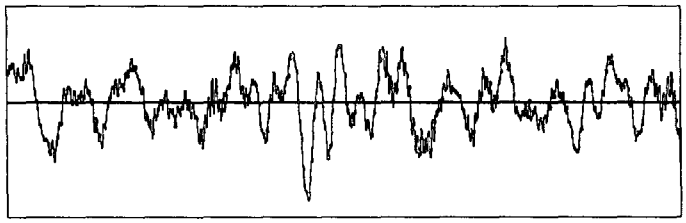
여기서, kmin과 kmax는 각각 사람의 피치 주기가 존재할 수 있는 최저 위치와 최고 위치를 나타낸다.

두 번째는 주기성 정도 V(n)이 연속되는 프레임들에 대해 얼마나 일정한 지를 나타내는 파라미터이다. 이는 후두암 환자의 음성이 개별 프레임의 주기성도 떨어지지만 여러 프레임에 걸쳐서 그 주기성 정도의 일정성도 떨어짐을 감안한 것이다. 진다. 본 논문에서는 현재 프레임과 중복되지 않고 가장 가까이 인접한 좌우 각각의 두 프레임과 현재 프레임을 포함하여 다섯 프레임에 대해 V(n)의 분산을 주기성의 동요 정도에 대한 파라미터 VAR_V(n)로 정의하였다. 만약 분석되는 프레임들 사이의 중복 구간이 없다면 VAR_V(n)은 다음 식과 같이 구할 수 있다.

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

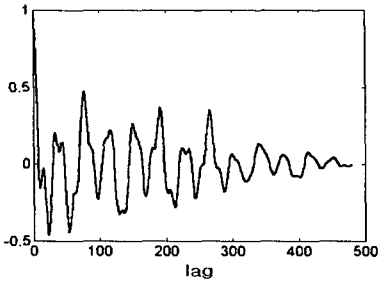


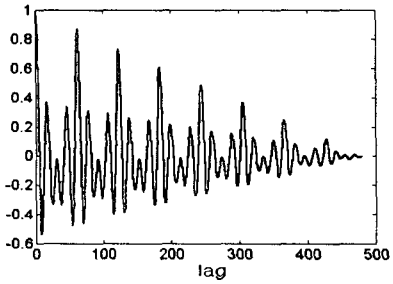
(a)



(b)

<그림 2> 정상인과 후두암 환자의 음성 파형의 예 (모음 /아/)
(a) 정상인의 음성 (b) 후두암 환자의 음성

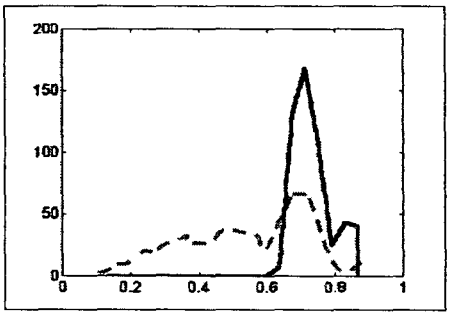


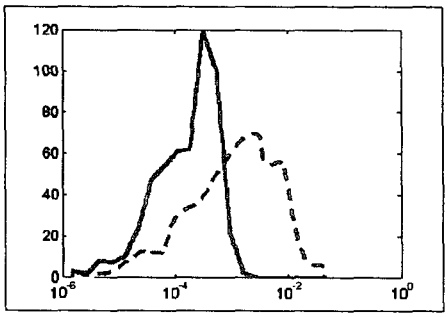


(a) (b)

<그림 3> 정상인과 후두암 환자의 정규화된 자기상관함수 Rn(k)/Rn(0)의 예
(a) 정상인의 경우 (b) 후두암 환자의 경우

<그림 4>는 임의의 정상인 26명과 임의의 후두암 환자 26명에 대한 주기성 정보 파라미터 값들의 분포를 보여 준다. 그림에서 정상인과 후두암 환자에 대해 주기성 특성에 차이가 있음을 확인할 수 있다. 그러나 이들 분포에 중복 구간이 존재하므로 주기성 특성만으로는 신뢰도 높은 식별이 가능하지 않음도 알 수 있다.





(a) (b)

<그림 4> 임의의 정상인(실선) 26명과 후두암 환자(점선) 26명에 대한 주기성 파라미터 값의 분포. (a) V(n) 값의 분포 (b) VAR_V(n) 값의 분포(가로축은 로그 스케일로 표현)

3. 후두질환 식별을 위한 인식기의 구성

본 논문에서는 후두질환 식별을 위한 인식기로 신경회로망과 Gaussian Mixture Model(GMM) 인식기를 사용하였다.

3.1. 신경회로망 인식기

신경회로망 인식기를 위해 은닉층을 하나 가지고 있는 다층 퍼셉트론 신경회로망을 사용하였고, 신경회로망의 학습 규칙은 일반적인 오차 역전파 알고리즘을 이용하였다[10]. 신경회로망의 은닉층 노드 수는 10개, 출력층 노드수는 1개로 하였으며, 입력층의 노드는 특징 파라미터의 차원수와 동일하도록 했다. 입력층의 각 노드의 입력값들은 평균이 0, 분산이 1이 되도록 정규화하였다. 1차 식별은 음성의 각 분석 프레임마다 이루어지며, 한 사람에 대한 최종 판정은 전체 프레임의 절반 이상의 프레임이 정상으로 판정되면 그 사람의 음성은 정상인의 음성으로, 그 반대의 경우에는 그 사람의 음성은 후두암 환자의 음성으로 판정하였다.

3.2. GMM 인식기

본 논문에서는 신경망 인식기와 더불어 Gaussian Mixture Model(GMM) 기반의 인식기를 사용하였다. GMM은 다음 식과 같이 복수개의 Gaussian 확률분포들의 가중합으로 구성된다.

$$p(\vec{x} \mid \lambda) = \sum_{i}^{M} p_i b_i(\vec{x}) \qquad (3)$$

여기서 $\vec{x}$는 정상인과 환자의 음성 데이터로부터 추출한 특징 벡터이고, p_i는 i 번째 mixture의 가중치이고, M은 mixture의 개수이다. 그리고 $b_i(\vec{x})$는 모델 λ의 i 번째 mixture의 Gaussian 확률분포이다.

GMM을 훈련하기 위해 정상인과 후두암 환자의 훈련 데이터로 먼저 단일 mixture를 가지는 모델을 만들고, 원하는 mixture 개수가 만들어질 때까지 모델의 mixture 개수를 하나씩 증가시켜 가며 훈련 과정을 반복하였다. 본 논문에서는 훈련용 데이터의 크기를 고려하여 mixture의 수는 1에서 10까지의 값을 사용하였다.

4. 실험 방법 및 결과

4.1. 음성 데이터와 실험 방법

후두암 식별 실험을 위해 한국 장애음성 데이터베이스[11] 및 추가적으로 구축된 음성 데이터를 사용하였으며, 정상인 음성 50개, 후두암 환자 음성 105개를 사용하였다. 그 외에도 기존 장애 음성 분석 방법으로 널리 사용되는 MDVP(Multi-Dimensional Voice Program) 분석이 안 되는 후두암 환자 음성 28개도 실험에 포함시켰다. 음성 데이터로 단모음 /아/ 중에서 모음의 중앙 부분 20 프레임을 사용하였고 정상인과 후두암 환자의 모두 2/3의 화자들은 학습에, 나머지 1/3은 식별 시험에 사용하였다. 이 때, MDVP로 분석이 안되는 후두암 환자의 음성이 한쪽 데이터에 몰리지 않도록 이 데이터 역시 2/3가 학습 데이터에 1/3이 식별 시험 데이터에 들어가도록 하였다. 실험 결과의 일관성을 높이기 위해 학습 및 식별에 사용한 음성 데이터는 다수의 세트를 랜덤하게 선정하여 실험을 5번 수행한 이후 이들의 평균 식별 결과를 계산하였다. 음성 데이터의 샘플링 주파수는 16kHz이며, 양자화 비트 수는 16비트를 사용하였다. LPCC나 MFCC를 얻기 위해서 프레임에 Hamming 윈도우를 사용하였으며, pre-emphasis 계수는 0.97로 하였다.

신경회로망 및 GMM 인식기의 훈련은 프레임 단위로 수행하고, 인식 과정 역시 1차적으로 프레임 단위로 수행하고 최종적으로 발화 단위로 식별 결과를 내도록 하였다.

4.2. 실험 결과

먼저, 신경회로망 인식기와 GMM 인식기에서 MFCC와 LPCC 계수의 성능을 비교하였으며 그 결과가 <표 1>에 나타나 있다. 이 실험에서 MFCC와 LPCC 계수는 모두 동일하게 16차로 고정시켰으며, GMM 인식기의 경우 mixture의 수를 1에서 10까지 변화시켜 가면서 실험을 수행하였고, 표의 결과는 mixture의 수가 4개일 때의 결과이다. <표 1>을 볼 때, 전반적으로 LPCC가 MFCC보다 높은 인식률을 나타내었고, 신경회로망 인식기와 GMM 인식기의 성능 차이는 별로 없었다. 이후의 실험에서는 신경회로망 인식기만을 사용하였다.

<표 1> MFCC와 LPCC의 식별 성능 비교

(a) 신경회로망 인식기를 사용할 경우

	정상음성	후두암음성	
		MDVP 분석가능	MDVP 분석불가능
MFCC(16차)	72.9%	92.0%	99.9%
LPCC(16차)	88.2%	96.0%	96.0%

(b) GMM 인식기를 사용할 경우

	정상음성	후두암음성	
		MDVP 분석가능	MDVP 분석불가능
MFCC(16차)	78.7%	90.6%	100.0%
LPCC(16차)	88.7%	95.3%	97.8%

<표 2>는 LPCC 계수의 차이가 식별 성능에 미치는 영향을 살펴본 것이다. 낮은 차수에 비해서 LPCC의 차수가 더 높은 경우에 식별 성능이 상대적으로 우수하였다. 이는 동일한 모음 발음에 대해 스펙트럼 포락선 정보를 더 정교하게 표현함으로써 정상인의 음성과 후두암 환자의 음성을 더 잘 구별해 주기 때문으로 판단된다. 본 논문에서는 16차보다 더 높은 차수에 대해서는 고려하지 않았으나, 이에 대한 검토도 필요할 것으로 판단된다.

마지막으로 <표 3>은 주기성 파라미터의 인식률에 대한 기여도를 알아보기 위한 실험 결과이다. 실험 결과 켑스트럼 계수와 주기성 파라미터를 함께 사용함으로써 인식 성능이 향상되며, 이들 두 종류의 파라미터는 상호 보완적인 정보가 될 수 있음을 보여주고 있다.

<표 2> 신경회로망 인식기에서 LPCC 차수에 따른 식별 성능 비교

	정상음성	후두암음성	
		MDVP 분석가능	MDVP 분석불가능
LPCC(12차)	88.3%	95.4%	90.0%
LPCC(16차)	88.2%	96.0%	96.0%

<표 3> 신경회로망 인식기에서 LPCC에 주기성 파라미터 추가에 따른 식별 성능 비교

	정상음성	후두암음성	
		MDVP 분석가능	MDVP 분석불가능
LPCC(16차)	88.2%	96.0%	96.0%
LPCC(16차)+주기성	89.4%	98.3%	96.0%

정리하면 높은 차수의 LPCC 계수를 주기성 파라미터와 함께 사용할 때 가장 우수한 식별 성능을 얻을 수 있었다. 그리고, 모든 경우에서 기존의 MDVP로 분석이 안 되는 데이터에 대한 인식률이 90%이상의 인식률을 보여 본 논문에서 사용한 방법의 유용성을 확인할 수 있었다.

5. 결 론

본 논문에서는 음성인식 기술을 기반으로 한 후두질환 음성의 자동 식별 방식들을 검토하고 그 성능을 비교하였다. 특징 분석 방법으로는 MFCC와 LPCC의 두 가지 켑스트럼 계수를 사용하였고, 음성 스펙트럼의 포락선 정보로 표현할 수 없는 주기성 정보를 추출하기 위해 자기상관 함수 기반의 주기성 파라미터들을 도입하였다. 인식기로는 다층 퍼셉트론 신경회로망과 GMM 인식기를 사용하였다. 실험 결과 신경회로망 인식기와 GMM 인식기의 성능은 비슷했으며, MFCC보다 LPCC가 전반적으로 우수한 성능을 보였다. 그리고, 음성 스펙트럼의 세부적인 정보를 얻기 위해서는 높은 차수의 LPCC를 사용하는 것이 적절함을 확인하였다. 또한, LPCC와 보완적인 정보를 가지는 주기성 파라미터를 추가함으로써 켑스트럼 계수만을 단독으로 사용하는 것보다 더 나은 결과를 얻을 수 있었다.

본 논문에서는 제한된 음성 데이터에 의한 신뢰도 문제의 보완을 위해 다양한 훈련 및 인식용 데이터의 조합에 의해 식별 결과를 얻었지만, 향후 보다 신뢰도 높은 평가를 위해 장애 음성 데이터베이스의 확충이 필수적으로 요청된다. 이와 더불어 성능 향상을 위해 다양한 특징 파라미터에 대한 추가적인 검토도 병행되어야 할 것이다.

참 고 문 헌

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

[2] S. Iwata, "Periodicities of pitch perturbation in normal and pathologic larynxes", *Laryngoscope*, Vol. 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.*, Vol. 71, pp.1544-1550, 1982.

[4] H. Kasuya, S. Ogawa et al., "Normalized noise energy as an acoustic measure to evaluate pathologic voice", *J. Acoust. Soc. Am.*, Vol. 80, No. 5, 1986.

[5] Y. Koike, H. Takhashi, T. C. Calcatera, "Acoustic measurements for detecting laryngeal pathology", *Acta Otolaryngol*, Vol. 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 석사학위 논문, 2002.

[9] B. Juang, L. Rabiner, *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, 창원대학교, 1999.

접수일자: 2003년 2월 10일
게재결정: 2003년 3월 8일

▶ 강현민(Hyun Min Kang)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과
소속: 부산대학교 전자공학과 지능정보처리연구실
전화: 051) 510-1698
FAX: 051) 515-5190
E-mail: kanghm@pusan.ac.kr

▶ 김수미(Soo Mi Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-1704
FAX: 051) 515-5190
E-mail: noise2@pusan.ac.kr

▶ 김유신(Yoo Shin Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과
소속: 부산대학교 전자공학과 지능정보처리연구실
전화: 051) 510-2376
FAX: 051) 515-5190
E-mail: kimys@pusan.ac.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
FAX: 051) 515-5190
E-mail: kimhs@pusan.ac.kr

▶ 조철우(Cheol-Woo Jo)
주소: 641-773 경남 창원시 사림동 9 창원대학교 제어계측공학과
소속: 창원대학교 제어계측공학과
전화: 055) 279-7552
E-mail: cwjo@sarim.changwon.ac.kr

▶ 양병곤(Cheol-Woo Jo)
주소: 614-714 부산시 부산진구 가야동 24 동의대학교 영어학과
소속: 동의대학교 영어학과
전화: 051) 890-1227
FAX: 051) 890-1222
E-mail: bgyang@dongeui.ac.kr

▶ 왕수건(Soo-Geun Wang)
주소: 602-739 부산시 서구 아미동 1-10 부산대학교 의과대학 이비인후과
소속: 부산대학교 의대 이비인후과
전화: 051) 240-7331
E-mail: wangsg@pusan.ac.kr