화자인식에서 차분을 이용한 새로운 데이터 추출 방법*

New Data Extraction Method using the Difference in Speaker Recognition

서 창 우**. 고 희 애**. 임 영 환**. 최 민 정***. 이 윤 정****

Changwoo Seo . Heeae Ko . Yonghwan Lim . Minjung Choi . Younjeong Lee

ABSTRACT

This paper proposes the method to extract new feature vectors using the difference between the cepstrum for static characteristics and delta cepstrum for dynamic characteristics in speaker recognition (SR). The difference vector (DV) which it proposes from this paper is containing the static and the dynamic characteristics simultaneously at the intermediate characteristic vector which uses the deference between the static and the dynamic characteristics and as the characteristic vector which is new there is a possibility of doing. Compared to the conventional method, the proposed method can achieve new feature vector without increasing of new parameter, but only need the calculation process for the difference between the cepstrum and delta cepstrum. Experimental results show that the proposed method has a good performance more than 2.03%, on average, compared with conventional method in speaker identification (SI).

Keywords: Speaker Recognition, Linear Prediction Cepstral Coefficient, Difference Vector, Gaussian Mixture models, Expectation-Maximization

* 본 연구는 2008 년도 숭실대학교 교내연구비 지원으로 이루어졌습니다. 그리고 논문의 저자들은 논문 심사기간 동안 심사위원님들의 정확하고 많은 지적에 감사하다는 뜻을 전하고 싶습니다.

** 숭실대학교 미디어학부

*** (주)인스모바일 기술연구소

**** 국방과학연구소

I. 서 론

사람의 음성신호 (speech signal)에는 음운정보뿐만 아니라, 각 개개인의 독특한 생체정보를 가지고 있다. 각 개개인의 생체적 정보로서의 음성신호를 이용하여 누구의 음성인지 알아내는 방법을 화자 인식(speaker recognition)이라고 한다[1].

관측된 음성신호로부터 효율적으로 화자의 특징벡터를 추출하는 방법에는 여러 종류가 있다. 현재 널리 사용되는 단구간 스펙트럼(short-term spectrum) 분석 방법인 LPCC(linear prediction cepstral coefficients)[2]는 음성 발생 모델인 전극 필터(all-poll filter) 모델로써 음성신호가 과거 몇 개의 신호들의 선형결합으로 예측할 수 있다고 가정한다. LPCC는 예측오차 제곱의 합을 최소값으로 하는 계수를 구하는 방법으로 자기상관(autocorrelation) 방법과 공분산(covariance) 방법이 있다. LPCC로 구해진 특징벡터는 정적 특성(static characteristic)만을 내포하기 때문에 시간정보를 갖는 동적 특성(dynamic characteristic)을 추가하기 위해서 델타 켑스트럼(delta cepstrum)을 적용하여 성능을 개선하였다[3]. 그러나 켑스트럼(cepstrum)과 델타 켑스트럼(delta cepstrum)이 정적 특성과 동적 특성을 각각 내포하고 있지만, 두 특성간의 차이에 대한 정보를 갖는 특징벡터는 없는 상태이다.

따라서 본 논문에서는 켑스트럼과 델타 켑스트럼의 차분벡터(DV: difference vector)를 이용한 정적 특성과 동적 특성의 중간적 의미의 새로운 특징벡터를 제안하였다. 새로운 특징벡터는 켑스트럼과 델타 켑스트럼의 차를 통해서 계산되기 때문에 정적 특성과 동적 특성의 중간적 의미를 내포하고 있으며, 새로운 파라미터의 추가 없이 적용할 수 있다.

논문은 다음과 같이 구성되었다. 2 장에서는 전처리 과정에서의 일반적인 특징벡터의 추출 방법에 대해서 설명하였다. 3 장은 논문에서 제안한 차분을 이용한 새로운 특징벡터에 대한 추출 방법을 설명하였다. 그리고 실험결과는 4 장에서 설명하였고, 5 장에서는 결론을 서술하였다.

2. 전처리 과정에서의 특징벡터

화자인식에서 단구간 스펙트럼을 이용한 특징벡터를 추출하기 위해서는 먼저 음성신호의 시작과 끝에서 불필요한 묵음구간을 제거하기 위해서 끝점 검출 (endpoint detection)을 수행해야만 한다[1]. 음성의 끝점을 검출한 후에 프리엠퍼시스 (pre-emphasis)를 수행하는데, 이는 고주파 포만트 (high frequency formant)와 저주파 포만트 (low frequency formant)의 진폭 (amplitude)을 비슷하게 얻기 위한 전처리 (preprocess) 과정이다. 프리엠퍼시스를 거쳐 전체 음성구간을 프레임 단위로 분리하는데, 일반적으로 화자인식 시스템에서는 스펙트럼 누출 현상을 최소화하기 위해 코사인 함수를 이용한 해밍 윈도우 (hamming window)나 해닝 윈도우 (hanning window)를 사용하여 50 % 겹침(overlap)으로 프레임을 생성한다. 다음 단계로 켑스트럼을 추출하기 위해서 LPCC 방법을 적용하고 있다[2].

이런 과정으로 계산된 켑스트럼은 채널의 영향을 줄이기 위해서 CMS(cepstral mean subtraction)를 사용하고 [4], 또한 고차성분을 강조하기 위해서 가중치(weighting) 함수를 적용함으로써 인식율을 높일 수 있다[5]. 그리고 켑스트럼을 이용한 유사한 특징벡터로는 델타 켑스트럼(delta cepstrum), 델타델타 켑스트럼(delta-dleta cepstrum) 등이 있다[3][6]. LPCC로부터 계산된 특징벡터는 정적 특성을 갖고 있는데 반해서, 델타와 델타델타 켑스트럼은 시간 미분 근사화(time-derivative approximations)를 통한 음성신호의 동적 특성을 갖는 특징벡터이다. 미분 근사화는 시간 t에서 켑스트럼 계수 O_t가 있을 때, 다음과 같이 델타 켑스트럼 계수를 구할 수 있다[6].

$$\Delta^i\{o_t\} = \frac{\sum_{\theta=1}^{\Theta}\theta\left(\Delta^{i-1}\{o_{t+\theta}\} - \Delta^{i-1}\{o_{t-\theta}\}\right)}{2\sum_{\theta=1}^{\Theta}\theta^2}, \qquad \Delta^0\{o_t\} = o_t \tag{1}$$

따라서, 켑스트럼 차수 k 에서 한 프레임에 대한 델타 켑스트럼은 다음과 같이 1 차와 2 차 미분 값으로 구성되어 있다.

$$\Delta o = \left[\Delta\{o(k)\},\ \Delta^2\{o(k)\}\right]. \tag{2}$$

그리고 델타 켑스트럼을 이용한 특징 파라미터의 일반적인 구성은 다음과 같이 요약할 수 있다.

$$O = \left[o(k),\ \Delta\{o(k)\},\ \Delta^2\{o(k)\}\right]. \tag{3}$$

3. 특징벡터간 차분을 이용한 새로운 특징값 추출

3 장에서는 화자인식에서 정적 특성을 갖고 있는 켑스트럼과 동적 특성을 갖는 델타 켑스트럼을 이용해서 새로운 특징 파라미터를 추출하는 방법을 설명하고자 한다. 새로운 특징벡터에 대한 생성 과정은 <그림 1>과 같이 입력신호로부터 2 장에서 소개된 켑스트럼과 델타 켑스트럼을 구한 후 각기 다른 특성을 갖는 두 특징벡터의 차를 계산함으로써 새로운 특징벡터를 구할 수 있다.

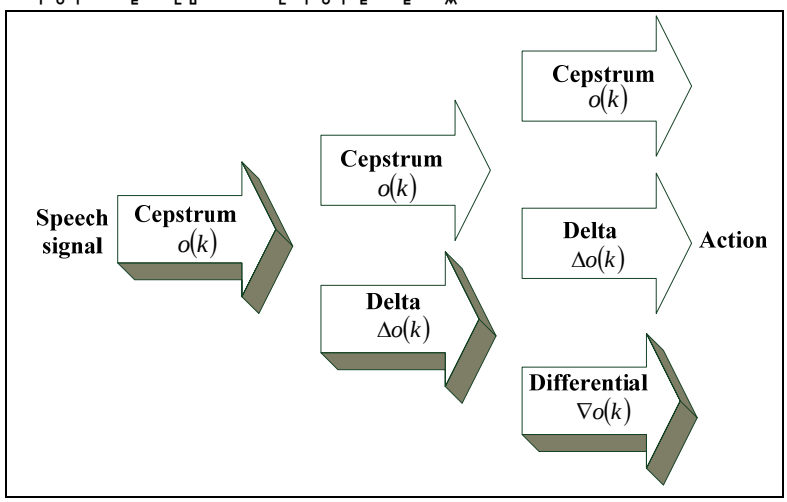

그림 1. 특징벡터의 생성 과정

이것을 수식적으로 나타내면 k 차수의 켑스트럼 $o(k)$와 델타 켑스트럼 $\Delta o(k)$의 차로 동적인 특성과 정적인 특성의 중간적 의미를 내포하고 있는 차분 벡터로 다음과 같이 나타낼 수 있다.

$$\nabla^i\{o(k)\} = \nabla^{i-1}\{o(k)\} - \Delta^i\{o(k)\}, \qquad \nabla^0\{o(k)\} = o(k) \tag{4}$$

따라서 이러한 차분 벡터는 1 차 차분 $\nabla o(k)$와 2 차 차분 $\nabla^2 o(k)$로 다음과 같이 구성할 수 있다.

$$\nabla o = \left[\nabla\{o(k)\}, \nabla^2\{o(k)\}\right] \tag{5}$$

본 논문에서 제안된 이러한 방법을 이용할 경우 특징 파라미터는 식 (3)보다 다양한 형태의 특징 파라미터로 구성할 수 있다. 먼저 정적 특성 $o(k)$, 동적 특성 $\Delta o(k)$, 그리고 두 특성의 중간적 의미를 내포하는 $\nabla o(k)$ 형태로 구성할 경우 다음과 같이 특징벡터를 구성할 수 있다.

$$O_1 = \left[o(k), \Delta\{o(k)\}, \nabla\{o(k)\}\right] \tag{6}$$

또한 시스템 구성에서 동적 특성이 강한 시스템으로 접근할 경우에는 정적 특성의 켑스트럼을 배제한 델타 켑스트럼과 1 차와 2 차 차분 벡터로 다음과 같이 구성할 수 있다.

$$O_2 = \left[\Delta\{o(k)\}, \nabla^i\{o(k)\}\right] \tag{7}$$

마지막으로 특징벡터 구성에서 정적 특성이 강한 시스템을 구성할 때 동적 특성의 델타 켑스트럼 대신 정적 특성이 강한 켑스트럼과 1 차와 2 차 차분 벡터로 구성할 수 있다.

$$O_3 = \left[o(k), \nabla^i\{o(k)\}\right] \tag{8}$$

4. 실험 결과

실험을 위해서 사용된 데이터는 대학원 실험실 환경에서 수집하였으며, 한국어 문장 종속 연속음(text-dependent continuous-speech)인 "무궁화 꽃이 피었습니다"를 사용하였다. 수집된 데이터는 1주 간격의 시간차를 가지고 있으며, 3주에 걸쳐서 수집하였다. 매주 1회 발성에서는 각 5 번 발성을 하였으며, 개인별 전체 발성된 데이터 수는 15 개이다. 수집된 데이터의 화자 인원수는 200 명으로 남/여 각각 100 명이다. 따라서 수집된 전체 데이터는 3,000 개이며 샘플링 주파수(sampling frequency)는 11.025 kHz이고 분해능은 16 bit이다. 개인별 학습을 위

해서 사용한 데이터는 처음 2 주간 10 번 발성한 데이터를 사용하였고 테스트 데이터는 마지막 주에 5 번 발성한 데이터를 사용하였다.

먼저 실험에서 사용된 음성의 프레임 길이는 180 샘플을 한 프레임으로 간주하고 프레임간 50% 중첩을 적용하였다. 특징벡터 추출을 위한 전처리(preprocessing) 과정에서는 프리엠퍼시스(pre-emphasis)와 해밍 윈도우(hamming window)를 사용하였으며, LPCC를 통해서 12차(order) 켑스트럼을 적용하였다. 또한 켑스트럼에서 채널의 영향을 줄이기 위해서 CMS[4]를 적용하고 식 (1)의 델타 켑스트럼에서 사용한 Θ 는 2를 적용하였다.

실험에서 화자의 특징 파라미터를 모델링하기 위해서 GMM(Gaussian Mixture models)을 사용하였고 [7], GMM의 학습 파라미터 추정은 EM(Expectation-Maximization) 알고리즘을 사용하였다[8]. EM 알고리즘의 기본적인 방법은 초기모델 λ로 시작해서 새로운 모델 $\hat{\lambda}$, 즉 확률밀도함수 $p(x|\hat{\lambda}) \geq p(x|\lambda)$를 추정하는 것이다. 또한 GMM에 사용된 파라미터 λ의 초기치는 VQ(vector quantization)를 이용해서 구하였다[9].

본 논문에서 제안한 방법이 우수하다는 것을 확인하기 위해서 수행한 비교 실험은 화자인식(SR : speaker recognition)방법 중에서도 가장 널리 수행되고 있는 화자식별(SI : speaker identification)을 적용하였다[7]. 화자식별(SI)은 N명의 이미 알고 있는 화자의 그룹으로부터 발성한 화자를 찾아내는 것이다. 이 방법은 각각의 화자의 특징을 테스트 화자의 특징과 매칭시켜서 가장 높은 유사도(likelihood)를 갖는 화자를 찾아내는 것이다. 또한 실험의 모델링을 위해서는 GMM의 Mixture 개수를 증가시키면서 기존의 방법과 제안한 방법과의 성능을 비교하는 방법으로 접근하였다.

먼저 첫 번째실험을 통한 제안한 방법의 우수성을 확인하기 위해서 비교한 방법은 12 차의 일반적인 정적 특성인 LPCC, 동적 특성의 델타, 그리고 논문에서 제안한 차분 벡터를 <그림 2>와 같이 비교하였다. 실험 결과 제안한 DV는 정적 특성을 갖는 LPCC보다 약 2.35%가 향상되었다. 그리고 일반적으로 델타 켑스트럼을 단독으로 사용하지는 않지만 DV와 비교했을 때 제안한 방법과 31.1%의 성능 차이를 보였다. 이 실험을 통해서 알 수 있는 것은 정적 특성과 동적 특성을 갖는 각각의 특징벡터를 사용하는 것보다 정적 특성과 동적 특성의 중간적 의미를 내포하고 있는 제안한 방법에서 구해진 특징 파라미터 DV가 화자식별(SI)에 우수하다는 것을 알 수 있다.

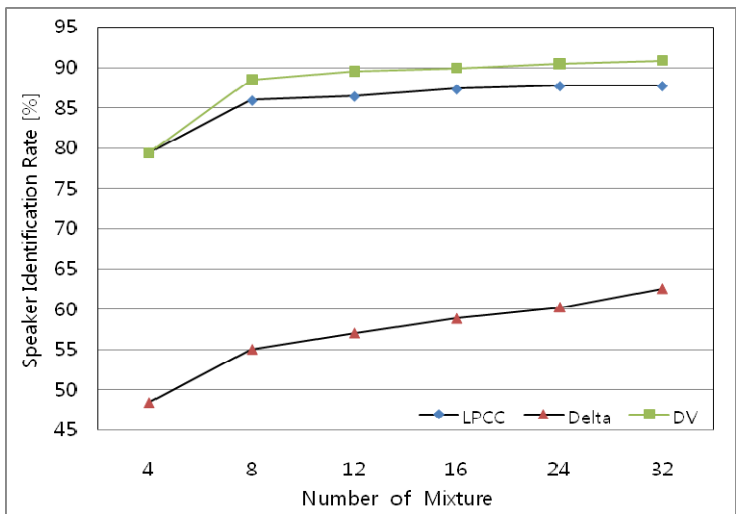

그림 2. 화자식별에서 각 특성별 일반적인 특징벡터(LPCC, Delta)와 제안한 방법(DV)의 성능 비교

<그림 3>은 LPCC에 1차 델타 켑스트럼을 적용한 24차의 일반적인 특징벡터의 구성인 LPCC/Delta에 제안한 차분 벡터인 DV를 LPCC와 Delta에 각각 조합한 DV/LPCC, DV/Delta를 비교하였다. <그림 3>의 실험 결과 제안한 DV/LPCC와 DV/Delta는 기존의 LPCC/Delta보다 0.5%와 1.93%가 개선되었다. 따라서 차분 벡터 DV를 정적인 특징벡터 LPCC와 함께 사용하는 것보다 동적 특성을 갖는 특징벡터인 델타와 함께 사용하는 것이 성능을 더욱 향상시키는 결과를 가져왔다. 이것은 DV가 두 특성의 중간적 의미를 내포하고 있지만, 실제로 정적인 특성이 강하기 때문에 동적인 특성의 Delta와 결합하는 것이 더욱 더 성능을 향상시키는 결과를 가져왔다.

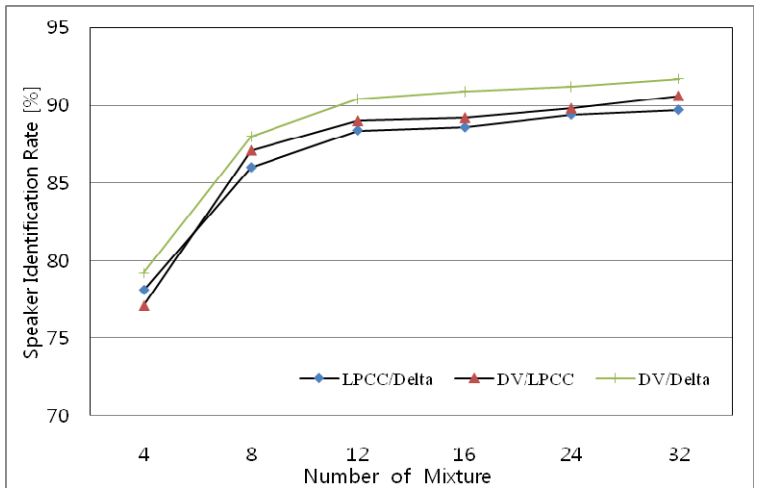

그림 3. 화자식별에서 일반적인 방법(LPCC/Delta)과 제안한 방법(DV/LPCC, DV/Delta)의 성능 비교

마지막으로 수행한 실험은 LPCC에 1차와 2차 델타 켑스트럼을 적용한 36차의 일반적인 특징벡터의 구성인 LPCC/Delta/Delta와 제안한 DV기반의 DV/LPCC/Delta, DV1/DV2/Delta, 그리고 DV1/DV2/LPCC를 비교하였다. 여기서 DV1과 DV2는 DV에 대한 1차와 2차를 나타낸다. 실험 결과 <그림 4>와 같이 제안한 방법의 DV/LPCC/Delta, DV1/DV2/Delta, 그리고 DV1/ DV2/LPCC는 각각 1.91%, 2.7%, 그리고 2.93%의 비교적 큰 폭의 성능 개선이 일어났으며, 평균 2.51%의 인식율이 향상되었다.

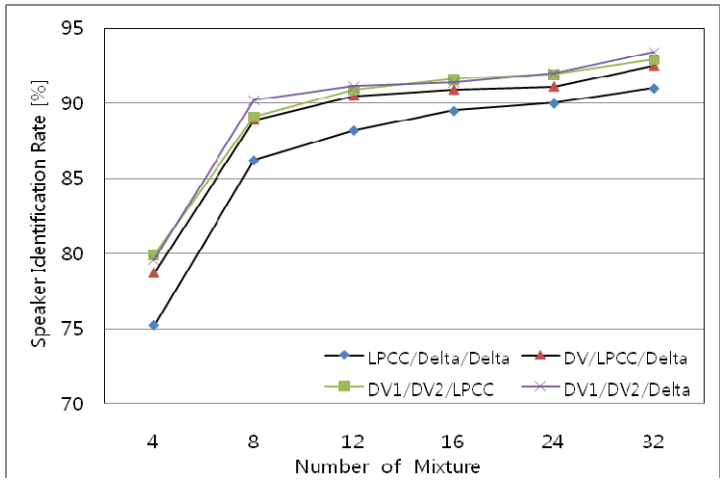

그림 4. 화자식별에서 일반적인 구성(LPCC/Delta/Delta)과 제안한 방법(DV/LPCC/Delta, DV1/DV2/ Delta, DV1/DV2/LPCC)의 성능 비교

따라서 실험 결과 본 논문에서 제안한 새로운 특징 파라미터는 <그림 2>에서 정적 특성의 켑스트럼과 비교 했을 때 평균 2.35%의 높은 성능 개선을 보였고, <그림 3>에서는 기존의 정적 특성과 동적 특성을 함께 사용한 시스템과 비교 했을 때 평균 1.22% 정도 향상되었다. 마지막 실험에서도 마찬가지로 기존의 특징벡터들로 구성된 방법과 비교 했을 때 보다 평균 2.51%의 인식율이 향상되었다. 실험 결과 본 논문에서 제안한 차분 벡터로부터 추출된 새로운 특징벡터를 기존의 특징벡터들로 구성된 것과 비교 했을 때 평균 2.03%의 인식율 향상을 가져왔다.

5. 결 론

본 논문에서는 화자식별(speaker verification)에서 정적 특성(static characteristic)을 내포하고 있는 켑스트럼(cepstrum)과 동적 특성(dynamic characteristic)을 내포하는 델타 켑스트럼(delta-cepstrum) 간의 차를 이용한 새로운 특징 파라미터를 추출하는 방법을 제안하였다. 켑스트럼과 델타 켑스트럼은 정적인 의미와 동적인 의미를 각각 내포하는 특징벡터로서 화자인식에서 널리 사용되어온 효과적인 알고리즘이다. 논문에서 제안한 차분 벡터(difference vector)는 정적 특성과 동적 특성의 차를 이용한 중간적 특징 값으로 새로운 특징벡터라고 할 수 있다. 따라서 제안한 차분 벡터는 기존의 특징벡터에서 켑스트럼과 델타 켑스트럼 간의 차를 구하는 과정이 추가되지만, 파라미터 개수가 증가되는 것은 아니다. 실험 결과 제안한 방법은 기존의 방법보다 화자식별에서 평균 2.03%의 높은 인식율 향상을 보였다.

참고 문헌

[1] Rabiner, L. R. & Juang, B. H. 1993. *Fundamental of Speech Recognition.* Englewood Cliffs, New Jersey: Prentice-Hall.

[2] Rothenberg, M. 1979. "A new inverse-filtering technique for deriving the glottal air flow wave and deriving voicing." *Journal of Acoustic Societyal America* 5, 1632-1645.

[3] Furui, S. 1986. "Speaker independent isolated word recognition using dynamic features of speech spectrum." *IEEE Trans. Acoustics Speech Signal Processing* 34, 52-59.

[4] Naik, D. 1995. "Pole-filtered cepstral mean subtraction." in *Proc. ICASSP-1995* 1, 157-160.

[5] Juang, B. H., Rabiner, L. R. & Wilpon, J. G. 1987. "On the use of Bandpass Liftering in Speech Recognition." *IEEE Trans. on ASSP* 35(7), 947-953.

[6] Young, S. 2001. The HTK Book, Cambridge University.

[7] Reynolds, D. A. & Rose, R. C. 1995. "Robust text-independent speaker identification using Gaussian mixture speaker models." *IEEE Trans, Speech Audio Proc.* 3(1), 72-82.

[8] Baum, L. E., Petrie, C. S. & Weiss, N. 1970. "Maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." *Ann. Math, Statist.* 41, 164-171.

[9] Lind, Y., Buzo, A. & Gray, R. M. 1980. "An algorithm for vector quantizer design." *IEEE Trans. Commun.* 28, 84-95.

접수 일자: 2008. 7. 23
수정일자: 2008. 8. 27
게재결정: 2008. 9. 4

▲ 서창우
서울시 동작구 상도동 511번지 (우: 156-743)
숭실대학교 미디어학부
Tel: +82-2-826-9872
E-mail: cwseo@ssu.ac.kr

▲ 고희애
서울시 동작구 상도동 511번지 (우: 156-743)
숭실대학교 미디어학부
Tel: +82-2-826-9872
E-mail: heeae@uniwebs.co.kr

▲ 임영환
서울시 동작구 상도동 511번지 (우: 156-743)
숭실대학교 미디어학부
Tel: +82-2-820-0685
E-mail: yhlim@ssu.ac.kr

▲ 최민정
경기도 성남시 분당구 야탑동 342-1 리더스 B/D (우: 463-828)
(주)인스모바일 기술연구소
Tel: +82-31-703-7301
E-mail: cmj1109@korea.com

▲ 이윤정
서울시 송파구 거여동 산25번지 (우: 138-110)
국방과학연구소
Tel: +82-2-3400-2684
E-mail: youn@add.re.kr