음성 인식률 개선방법에 관한 연구

김영포* · 이한영**

A Study on Improved Method of Voice Recognition Rate

Young-Po Kim* · Han-Young Lee**

요 약

본 논문에서는 음성 인식률 개선에 관한 방법을 제시하고 연구하였다. 기존의 음성 검출 방법 중 많이 이용되고 있는 HMM(Hidden Markov Model) 알고리즘을 이용하여서 음성을 검출하였다. 실험은 음성 검출과 음성 인식의 두 가지 방법으로 진행하였다. 음성 검출은 음성의 단위로 영교차율을 구하여 데이터의 유무를 판별하였다. 음성 인식은 음성의 형상의 패턴을 분석한 후 학습된 패턴과 비교 하는 형식으로 분석하였다. 실험 결과, 제안된 음성 형상의 패턴인식 이용한 알고리즘은 92%의 음성 인식률을 얻어 80%의 기존 HMM 알고리즘에 비해서 약 12%의 향상된 인식률을 얻을 수 있었다.

ABSTRACT

In this paper, we suggested a method about the improvement of the voice recognition rate and carried out a study on it. In general, voices were detected by applying the most widely-used method, HMM (Hidden Markov Model) algorithm. Regarding the method of detecting voices, the zero crossing ratio was calculated based on the units of voices before the existence of data was identified. Regarding the method of recognizing voices, the patterns shown by the forms of voices were analyzed before they were compared to the patterns which had already been learned. According to the results of the experiment, in comparison with the recognition rate of 80% shown by the existing HMM algorithm, the suggested algorithm based on the recognition of the patterns shown by the forms of voices showed the recognition rate of 92%, reflecting the recognition rate improved by about 12% compared to the existing one.

키워드

Voice signal, Voice detecting, algorithm, Voice recognition
음성신호, 음성 검출, 알고리즘, 음성인식

Ⅰ. 서 론

음성은 인간이 가지고 있는 기본적인 능력 중에서 가장 중요한 것으로 기본적인 의사소통을 위한 수단이며, 편리함과 경제성의 측면에서 우수한 특징을 가진다. 또 음성에 의해 표현되는 말은 인간과 인간 사이의 의사소통의 수단뿐만 아니라 논리적으로 사물을 생각하는 경우에서도 중요한 역할을 한다. 일반적으로 음성 인식(Voice Recognition)이란 마이크나 전화기로 화자에 의해서 발성된 음향적인 신호를 인간이 이해할 수 있는 단어나 구문들로 표현하는 일련의 과정을 말하며, 최종적으로 인식된 단어나 구문을 컴퓨터나

* 한국항공대학교 정보통신공학과 박사과정(zeropo@kau.ac.kr)
** 대림대학교 방송음향영상학과(hylee@daelim.ac.kr) 교신저자
접수일자 : 2012. 09. 10 심사(수정)일자 : 2013. 01. 10 게재확정일자 : 2013. 01. 21

기계 상의 명령이나 제어, 자료입력, 그리고 문서의 준비 등을 위한 용도로써 이용되고 있다[1, 2]. 음성의 특징 파라미터를 추출하여 컴퓨터로 하여금 인지하도록 하는 것이다. 이미지를 이용한 그래픽 인터페이스의 발달로 마우스와 키보드를 병용함으로써 편의성이 많이 개선되었으나 지구에서 가장 오래되고 많이 사용하는 사람끼리의 대화에 비하면 불편한 점이 많다. 음성 인식기술은 이러한 휴먼 인터페이스를 편리하고 자연스럽도록 개선할 핵심기술 중 하나이다[3, 4].

본 논문에서는 기존의 LPC(Linear Predictive Cod- ing), MFCC(Mel Frequency Cepstral Coeffici- ents), VQ(Vector Quantization), HMM(Hidden Markov Model) 방법 중 많이 사용하고 있는 HMM 방법과 본 논문에서 제안한 형상의 패턴인식 알고리즘의 인식 성능을 비교·검토하였다.

II. 음성 인식 시스템 및 추출 알고리즘

2.1. 음성 인식 시스템

음성 인식 시스템은 인식의 대상에 따라 화자독립 시스템과 화자종속시스템 그리고 발음의 형태에 따라 고립어 인식 시스템과 연속어 인식 시스템으로 나눈다. 화자종속 음성 인식은 화자독립 음성 인식에 비해 인식률이 높아 실용화에 유리하며 대체로 화자종속 시스템의 성능이 화자독립의 시스템보다 인식률이 높게 나온다. 화자종속 시스템은 현재 휴대폰에 탑재되어 사용되는 음성 다이얼링 시스템에 이용되고 있다.

사용자 음성을 저장, 등록하여 실제 인식을 수행할 때는 입력된 음성의 패턴과 저장된 음성의 패턴을 비교하는 기법이다. 전화를 걸 때 사람 이름만 말하면, 전화번호를 찾아 자동으로 전화를 걸어주는데 이용되고 있다. 화자독립 시스템은 불특정 다수 화자의 음성을 인식하기 위한 시스템으로 시스템 동작 전 음성 등록의 번거로움이 없고 다수화자의 음성을 수집하여 통계적인 모델을 학습시키고, 학습된 모델을 이용하여 인식을 수행하여 각 화자의 특징적인 특성은 사라지고 각 화자 간에 공통으로 나타나는 특성이 부각하는 방법이다. 고립어 인식 시스템은 짧은 음성명령이나 간단한 음성제어 등에 주로 사용되며 숫자 음을 인식하여 음성버튼으로 사용하는 경우“1,2,3”과 같은 숫자보다 “일,이,삼”과 같은 말로 각 단어가 또박또박 발음되고 각 단어 사이에 충분한 길이의 묵음구간이 존재하여야 한다. 인식률 높고 구현하기 간단해 널리 이용되고 있으나 사용자 이용하기가 불편하다는 단점이 있다. 연속어 인식 시스템은 문장 단위로 인식을 수행하는 시스템으로 문장을 인식하기 때문에 사용자가 단어 단위로 끊어 발음하지 않아도 된다. 문장은 평상시와 같이 발음되며, 특별히 단어 사이의 묵음이 있을 필요는 없다. 연속어인 경우, 한 단어 특성이 인접한 단어의 발음에 영향을 받는 조음효과(Coarticulation Effect)는 연속어 인식을 어렵게 한다[5].

2.2. 음성 추출 알고리즘

음성추출 방법으로는 LPC, MFCC, VQ, HMM 등 여러 가지 음성추출 알고리즘이 있다. LPC(Linear Predictive Coding) 추출은 과거의 일정 개수의 샘플 값들에 계수를 각각 곱하고 이를 총합한 값으로 현재의 샘플 값을 예측하려는 시도에서 출발하였다. 계수는 선형예측계수(LPC)라 하고 전달함수 입장에서 보았을 때 전극(All-pole)모델을 이루며 LPC를 추출하는 과정으로는 구간 내 자기 상관계수를 구하고 이를 재귀적인 방법을 통해서 빠르게 계산을 하여 LPC 계수를 기반으로 하여 음성인식에 효과적인 캡스트럼 계수(LPCC)로 변환하여 사용한다[6].

MFCC(Mel Frequency Cepstral Coefficients) 추출 방법은 사람의 귀가 주파수 변화에 반응하게 되는 양상이 선형적이지 않고 멜 스케일을 따르는 청각적 특성 반영한 캡스트럼 계수 추출 방법이다. 멜 스케일에 따르면 낮은 주파수에서 작은 변화에 민감히 반응하지만, 높은 주파수로 갈수록 민감도가 작아지므로 특징 추출 시에 주파수 분석 빈도를 특성에 맞추는 방식이다.

VQ(Vector Quantization)방식은 연속 혹은 떨어진 벡터들을 코드 북과 맵핑 시켜 적절한 디지털 시퀀스로 부호화하는 방법이다. VQ의 목적은 데이터 감축으로 데이터의 충실도를 잃지 않으면서 비트율을 감소시키며 스칼라 대신 벡터 코딩 방식을 사용하는 것은 데이터의 양을 줄일 수 있어 스칼라 대신 벡터로 조합된 신호를 코딩하는 것이 적은 데이터율로 좋은 성능을 얻을 수 있다[7].

HMM(Hidden Markov Model)은 통계적 패턴인식

을 이용한 방법으로 통계적인 정보를 확률모델 형태로 저장하고 미지의 입력패턴이 들어오면 각각의 모델에서 이 패턴이 나올 수 있는 확률 계산한다. 음성신호를 상태천이 확률과 각 상태에서의 관찰확률 이라는 두 단계에 걸친 확률 과정으로 표현한다. 현재 가장 널리 사용되는 방법으로 음성처리 및 언어처리를 단일구조로 처리할 수 있다는 장점이 있다[8].

III. 음성 검출 알고리즘

3.1. 음성 획득

음성 분석 및 합성, 음성인식, 음성 부호화 등 음성신호처리의 거의 모든 분야에서 음성신호의 음성구간, 즉 음성의 시작점 및 끝점을 정확하게 추출해내는 일은 매우 중요하다. 디지털 신호의 처리에서 먼저 논해야 할 것은 바로 처리 대상의 신호가 '있다' 와 '없다'를 가려주는 것이다. 있지도 않은 신호 정보를 대상으로 기다리는 시간인 '공회전' 상태가 지속한다면 효율성 면에서 상당히 문제가 될 수 있다. 음성과 비 음성 신호 간의 경계점(End-Point) 검출은 반드시 선결되어야 하는 과제이며, 음성인식의 효율성에 직접적인 영향을 주게 된다. 음성신호를 부호화하는 과정에서 음성 활성도를 검출하여 음성구간만 부호화함으로써 전송효율을 높여 시스템의 용량을 증가시키는 효과를 가져 오게 된다. 음성신호에서 음성구간의 시작점과 끝점은 사람들의 발성 시에 만드는 인위적 결과(Artifact)에 따라 데이터와 잡음의 구별을 어렵게 한다. 정상적으로 발음된 음성구간도 그 음성의 특성에 따라 소리가 작은 부분, 큰 부분 또는 무성음인 부분, 유성음인 부분의 문제로 각각 다르게 나타난다. 따라서 잡음이 있는 음성신호에서 정확한 음성구간을 검출한다는 것은 쉽지 않은 일이다. 음성신호에서의 잡음으로는 배경잡음과 같은 부가잡음, 통신채널의 특성과 같은 콘볼루션 잡음 등을 들 수 있다. 이상적인 조건이 현실적으로 실현되기 어려워서 잡음과 음성의 분류에 대한 대책이 필요하다. 음성검출을 하기 위해 Rabinar와 Sambur의 에너지와 영교차를 이용한 음성검출 알고리즘, Lamel의 레벨 등화기(level equalizer)를 이용한 음성 검출 알고리즘, Teager의 에너지를 이용하는 방법 등의 알고리즘이 있다[9].

본 논문에서는 영교차를 이용하여 음성 신호의 시작과 끝점을 구별하여 음성 데이터의 유무를 판별하고 음성 형상의 패턴을 분석하는 알고리즘을 이용하여 음성을 인식하였다.

3.2. 개발 환경

음성 인식 시스템은 학습 과정과 검사 과정으로 구성되어있다. 학습 과정에서는 검사 대상 패턴의 위치와 속성을 학습하여 저장한다. 검사 과정에서는 검사 대상의 음성을 입력 받아 검사 할 패턴이 있는 부분만 추출하고 히스토그램을 분석하여 학습된 패턴과 비교하여 검사한다. 개발 환경은 펜티엄 듀얼 E5200과 마이크로폰을 사용하였고 개발 언어는 Visual C 6.0 버전이며. 음성 데이터는 CD 음질을 기준으로 44KHz, 16bit, 스테레오로 설정하였다. 완성된 프로그램은 그림 1과 같다.

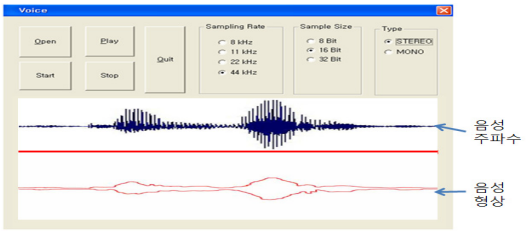

그림 1. 음성 인식 시스템
Fig. 1 Voice recognition system

3.3. 음성 검출

음성인식에서 S/N비가 30dB 이상으로 클 때 작은 에너지를 갖는 음성신호라도 주변잡음보다 큰 에너지 값을 가지므로 에너지 값을 이용하여 간단하게 음성구간을 검출할 수 있다. 하지만 이상적인 조건이 실현되기는 현실적으로 어렵기 때문에 잡음에 대한 대책이 필요하다. 음성검출은 분석구간 단위로 검출하므로 대상구간에서 음성 데이터를 구하여 신호의 여부를 판별한다. 한 분석 구간의 절대 에너지는 식 (1)와 같이 정의되고, 식(2)는 영교차율을 나타내며, 음성 구간은 그림 2와 같이 적용된다[10]. 음성 데이터의 상한 값과 하안 값을 정하여 일정 구간의 상한 값을 넘으면 구간에 음성이 있다고 간주한다. 유성음은 값이 크기 때문에 대부분 데이터가 검출된다. 하안 값보다 데이터의 값이 크면 바로 이전 구간이나 바로 이후 구

간이 유 음성 구간이면 유 음성으로, 바로 이전 구간이나 바로 이후 구간이 무 음성 구간이면 무 음성 구간으로 간주한다. 물론 에너지 하한 값 보다 작으면 무 음성으로 간주한다. 적합한 상한 값과 하안 값은 음성데이터의 수치 값에 따라 달라지므로 환경에 따라 자동으로 설정할 수 있게 구현하였다[11].

$$E=\frac{1}{N}\sum_{i=0}^{N-1}|x^2(i)|\ E_n(db)=10\log E_n \quad (1)$$

$$ZCR=\sum_{i=0}^{N-1}|sgnx(i)-sgn(x(i+1)|*(1/2) \quad (2)$$

$$sbn(x)= \quad 1\ (x\geq 0)$$
$$sbn(x)=-1\ (x<0)$$

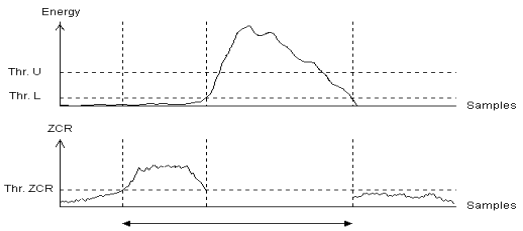

그림 2. 음성신호의 음성구간 특성
Fig. 2 Voice interval characteristics of voice signal

절대 에너지에 의한 음성의 판별에서 유성음은 쉽게 찾아낼 수 있으나 무성음의 경우에는 무음성과 별로 차이나지 않은 정도의 절대에너지 값밖에 가지고 있지 않다. 따라서 다른 방법으로의 검출이 필요하다. 무성음이 무음성과는 뚜렷이 구분되면서 유성음보다 오히려 큰 수치를 가지는 측정값이 있는데 이것이 영교차율 이다 [12].

Ⅳ. 음성 인식 검토

음성 인식은 음성 주파수를 분석 하여 형상을 만들어 내는 작업이다[13]. 히스토그램 분석은 검사하려는 음성 주파수를 분석하여 높은 레벨 쪽에 있는 히스토그램 값들 중에 최대값이 되는 레벨을 찾고 같은 방법으로 검사 대상을 구성하는 낮은 쪽의 레벨 중에 최대값이 되는 레벨을 찾는다. 두 레벨 사이의 가장 최소값을 갖는 레벨의 값을 이어서 형상을 구한다. 하지만 형상의 값을 찾는 단계에서 많은 굴곡이 있어 경계 값을 찾는데 어려움이 있다. 이런 문제를 해결하기 위하여 그림 3과 같이 데이터의 굴곡에 따라 몇 단계씩 묶어서 새로운 데이터 형상을 생성하였고, 식 (3)을 적용하여 그림 4와 같이 실제 음성 주파수의 형상을 쉽게 찾아 낼 수 있다[13-14].

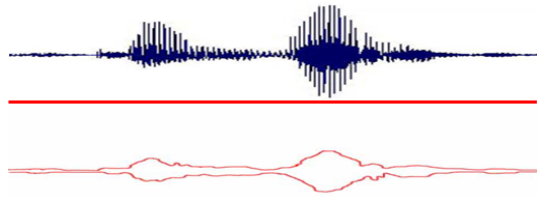

그림 3. 20dB S/N 음성 데이터와 검출된 형상패턴
Fig. 3 20dB S / N voice data and detected pattern

$$f(x)=\frac{f(x)+ + \bullet\ \bullet\ \bullet\ f(x+n)}{n+1} \quad (3)$$

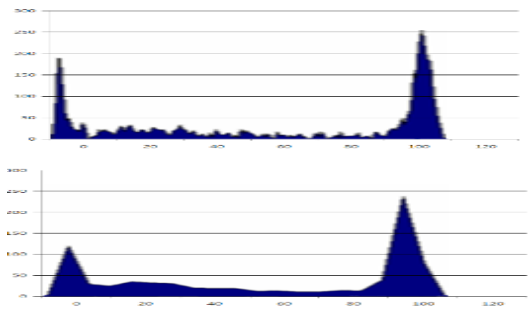

그림 4. 음성 주파수 데이터 형상
Fig. 4 Voice frequency data shape

찾아낸 형상과 음성 데이터간의 사이의 정확한 검사를 위해서는 두 데이터간의 좌표축을 정확히 일치시켜야 한다. 음성 데이터는 한쪽으로 직진하는 1차원 데이터 이므로 수평 좌표 보정을 필요로 한다[11]. 데이터의 축 고정을 위한 이동은 식 (4)와 같이 다항식 보간(polynomial interpolation) 함수를 음성 데이터 전체에 적용시켜 수행된다. $x'(x,y)$는 이동후의 새롭게 교정된 음성 주파수 좌표의 쌍 (x)를 나타내며 행렬 A 는 상관관계의 점(correlation points)으로부터 파생

된 다항식 이동(polynomial shift)의 계수이다.

$$x'(x,y) = X^T A Y \quad (4)$$

학습데이터는 5회분을 참조데이터로 생성하여 인식 평가를 하였다. 참조 데이터는 20dB 이하의 조용한 환경에서 획득한 음성 데이터를 사용하여 제작하였다.

일반적으로 음성 인식은 일상생활의 신호 대 잡음비를 고려하여 20dB이상을 요구한다. 본 실험에서는 각각의 S/N비를 5dB, 10dB, 15dB, 20dB로 분류하여 실험하였다. 전처리 과정으로 기존의 순수한 음성 데이터와 각 dB별 음성 데이터의 특징을 추출하여 음성 인식률을 높였다. 실험 방법은 HMM 알고리즘과 HMM 알고리즘에 형상 패턴인식 방법을 더한 알고리즘을 이용하여 각 dB별 인식률을 계산하였다.

실험결과 HMM 알고리즘보다 본 논문에서 제안한 알고리즘이 S/N비율이 5dB에서 2%의 개선을 보이며 인식률이 많이 떨어졌고 20dB일 경우 인식률이 11% 정도의 개선되었으며, 그 결과를 그림 5와 표 3에 나타내었다 [15-16].

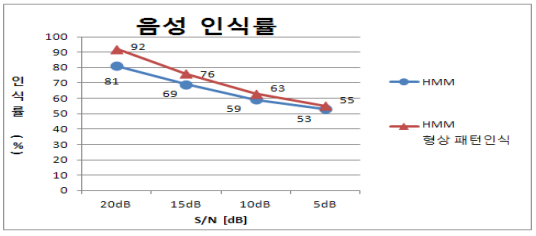

그림 5. 음성 인식률
Fig. 5 Voice recognition rate

표 1. 음성 인식표
Table 1. Voice recognition tags

알고리즘	S/N[dB]	인식률(%)
HMM	20	81
	15	69
	10	59
	5	53
HMM 형상 패턴인식	20	92
	15	76
	10	63
	5	55

기존 음성 데이터에 그림 6의 가우시안 노이즈를 추가하여 인식률을 다시 측정 하였다. 측정 결과 HMM방식은 음성 인식률이 10% 정도 저하되는 것을 보였지만, HMM 형상 패턴인식에서는 2~3%의 인식률의 저하를 보여 기존의 HMM방식에 비해 노이즈에 강한 인식률을 보여 HMM 형상 패턴인식 알고리즘이 노이즈에도 강하다는 것을 그림 7과 표 4에서 나타내었다 [17].

그림 6. 가우시안 노이즈
Fig. 6 Gaussian noise

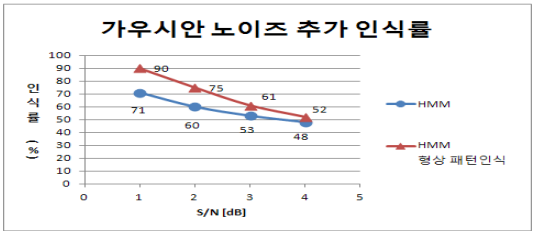

그림 7. 가우시안 노이즈 음성 인식률
Fig. 7 Voice recognition rate of Gaussian noise

표 2. 가우시안 노이즈 음성 인식표
Table 2. Voice recognition tags of Gaussian noise

알고리즘	S/N[dB]	인식률(%)
HMM + Gaussian noise	20	71
	15	60
	10	53
	5	48
HMM 형상 패턴인식 + Gaussian noise	20	90
	15	75
	10	61
	5	52

V. 결 론

본 논문에서는 음성 인식률 향상에 관한 방법을 제시하고 실험하였다. 잡음이 검출할 음성보다 크면 인식 성능이 떨어지고 사람조차도 음성을 인식하기가

어려운 잡음환경 상황이므로 본 실험에서는 잡음을 고려하지 않기로 하였다. 따라서 효과적인 음성 인식을 위해서 주변잡음 보다는 높은 음압레벨로 발성하였다. 본 논문에서 제안한 방법으로 실험한 결과 20dB의 S/N에서 기존의 HMM 알고리즘에 비해 약 11% 개선된 92%의 음성 인식률을 얻을 수 있었다.

특히, 주변에 일상적인 잡음이 있는 복도에서도 비교적 양호한 결과를 나타내었다. 그러나 데이터양이 많아질 경우 인식속도가 느려진다는 단점은 알고리즘의 효율성 문제로 남아있지만, 인식 단어의 수가 한정되면 문제없이 빠른 인식 속도를 보여주고 있다. 음성인식률을 개선시하기 위해서는 잡음환경을 잘 고려한 데이터 처리의 효과적인 모델화 방법과 음성패턴의 시간적 상관관계를 더욱 잘 표현 할 수 있는 전처리 과정에 대한 연구가 요구되며, 알고리즘에 대한 효율적이고 개선된 학습 알고리즘에 관한 연구가 필요하다. 본 논문에서 제안한 음성인식 알고리즘과 학습 DB를 접목하여 앞으로 잡음환경에서의 음성인식 시스템으로 확장할 수 있을 것으로 기대한다.

참고 문헌

[1] J. M. Markoul, “Linear prediction A tutorial review”, Proceeding sof IEEE, Vol. 63, No. 4, 1975.

[2] 齊藤-收三, 中田和男, “音聲情報處理の基礎”,オ-ム社, 1982.

[3] 中川聖一, “連ﾒ續出力分布型HMMによる日本 音韻認識”, 音響學會論文誌, Vol. 46, pp. 486-496, 1990.

[4] 허강인, “스펙트럼 모멘트법에의한 韓國語音聲의포만트周波數 推定에 관한 硏究”, 博士學位論文, 1990.

[5] 김수훈, 이종진, 허강인, “이산 지속시간제어 연속 분포 HMM을 이용한 연속음성 인식”, 한국음향학회논문지, 14권, 1호, pp. 81-89, 1995.

[6] 한학용, “우리말 음성의 최적분할과 인식에 관한 연구”, 博士學位論文, 2004.

[7] 한학용저, “패턴인식개론”, 한빛미디어, pp. 418-424, 2009.

[8] L. R. Rabiner and B. H. Juang, “An Introdu-ction to Hidden Markov Models”, IEEE ASSP magazine, pp. 4-17, 1986.

[9] 이종진, “한국어 연속음성 인식시스템의 구현”, 博士學位論文, 1994.

[10] 한학용저, “패턴인식개론”, 한빛미디어, pp. 418-424, 2009.

[11] A. Rosen Feld, A. C. Kak, “Digital Picture Processing”, 2nd Edition, Academic Press, 1982.

[12] Franco, H., Weintraub, M., Cohen, M., “Context modeling in a hybird HMM neural net speech recognition system”, International Conference on Neural Networks, Vol. 4, 9-12, pp. 2089-2092, 1997.

[13] 김용연, “영상통신을 위한 웨이블릿 변환 부호화 ”한국전자통신신학회논문지, 6권, 1호, pp. 61-67, 2011.

[14] 홍완표, “데이터 전송 효울을 wrhvy한 3x4비트 1 바이트 문자 부호화 규칙에 관한 연구” 한국전자통신신학회논문지, 6권, 4호, pp. 499-504, 2012.

[15] 이창영, “음성인식에서 중복성의 저감에 대한 연구”, 한국전자통신신학회논문지, 7권, 3호, pp. 475-483, 2012.

[16] 김범준 “버퍼 크기 기반 자동재전송 프로토콜의 재전송 지속성 제어” 한국전자통신신학회논문지, 7권, 3호, pp. 487-492, 2012.

[17] 정상래, “NCW 및 전술데이터링크 기술개발 현황분석”, 한국전자통신신학회논문지, 7권, 5호, pp. 991-998, 2012.

저자 소개

김영포(Young-Po Kim)

2001년 국립경상대학교 정보통신공학과 졸업(공학사)

2005년 한국항공대학교 대학원 항공전자공학과 졸업(공학석사)

2011년 한국항공대학교 대학원 정보통신공학과(박사수료)

※ 관심분야 : Mobile IP, Self-organizing wireless networks, 정보보안

이한영(Han-Young Lee)

1998년 세종대학교 물리학과 졸업(이학사)
2002년 국민대학교 대학원 전자공학과 졸업(공학석사)
2005년 건국대학교 대학원 전자정보통신학과 졸업(공학박사)
1999년~현재 대림대학교 방송음향영상과 학과행정기사
※ 관심분야 : 안테나, RFIC, Power amplifier, 음향시스템