韓國 컴퓨터情報學會 論文誌
第9卷 第4號, 2004. 12.

2004-9-4-1-12

에너지 라벨링 그룹화를 이용한 고속 음성인식시스템

한 수 영*, 김 홍 렬**, 이 기 희***

Fast Speech Recognition System using Classification of Energy Labeling

Su-Young Han*, Hong-Ryul Kim**, Kee-Hee Lee***

요 약

본 논문에서는 입력된 음성의 음소단위로 추출된 에너지 파라미터를 이용하여 에너지를 라벨링(energy labeling)하고 라벨링된 값에 따라 입력 음성을 그룹화하였다. 그리고 동적패턴정합 수행 시 입력된 실험 음성에서 검출된 에너지의 크기에 따라 선택된 라벨의 그룹 내에서 DTW를 수행시켜 처리시간을 단축시켜 저가형 프로세서에서도 고속으로 동작할 수 있게 하고자 하였다. 본 논문의 음성 라벨링 단계는 음성의 구간 검출 및 에너지 파라미터의 추출 단계에서 정확한 파라미터의 검출을 전제로 하기 때문에 이를 보완하기 위해 피치의 주기에 따른 가변윈도우를 사용하였다. 피치주기를 먼저 구하고 그 주기에 200 프레임에서 300프레임 사이에서 윈도우의 크기를 결정함으로써 윈도우의 영향이 제거된 에너지를 구하는 방법을 제안하였다. 실험결과 제안된 방법이 약 25% 정도의 연산량을 감소시켰다.

Abstract

In this paper, the Classification of Energy Labeling has been proposed. Energy parameters of input signal which are extracted from each phoneme are labelled. And groups of labelling according to detected energies of input signals are detected. Next, DTW processes in a selected group of labeling. This leads to DTW processing faster than a previous algorithm. In this Method, because an accurate detection of parameters is necessary on the assumption in steps of a detection of speeching duration and a detection of energy parameters, variable windows which are decided by pitch period are used. A pitch period is detected firstly; next window scale is decided between 200 frames and 300 frames. The proposed method makes it possible to cancel an influence of windows and reduces the computational complexity by 25%.

▸ Keyword : speech recognition, pattern classification, pitch detection

• 제1저자 : 한수영
• 접수일 : 2004.10.20, 심사완료일 : 2004.11.16
* 안양대학교 컴퓨터학과 전임강사, ** 동서울대학 컴퓨터정보과 조교수, *** 동서울대학 컴퓨터정보과 부교수

Ⅰ. 서 론

최근 음성과 자연언어의 기본적인 성질의 이해에 관한 관심이 높아지고 각종 디지털 미디어의 발달, 초고속 정보통신망 구축과 더불어 멀티미디어 통신을 통한 통신 판매, 물류처리, 제품홍보 등이 폭증하고 있으며 관공서 등에서도 대민 서비스 품질에 관한 관심도 점점 높아져가고 있다. 이와 더불어 개인용 컴퓨터 보급에 의한 신호처리기술과 정보처리기술의 급속한 발전과 함께 음성을 통한 인간과 기계와의 직접적인 커뮤니케이션을 위한 Man-Machine 인터페이스의 중요성도 강조되고 있다. 또 인간과 기계사이 뿐만 아니라 인간과 인간 사이에 기계를 통한 통역을 자동적으로 하고자 하는 연구도 활발히 진행되고 있다.

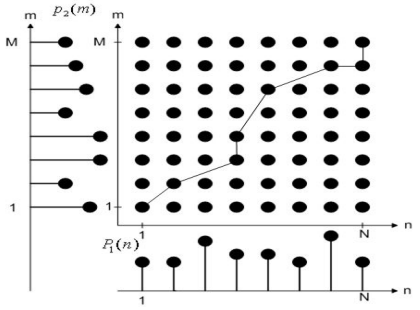
1960년대부터 음성 발성과 이해에 관해 많은 기초적 연구가 진행된 이래 기계에 의한 연속음성인식, 합성에는 아직 많은 과제가 남아있지만 최근 40여 년간 연구결과로 고립단어 인식에 많은 발전이 있어 미국, 유럽 일본 등에서는 상용제품도 출현하고 있다. 이들 인식시스템의 대부분은 고립단어, 또는 한정된 범주의 연속음성인식시스템이지만 잡음환경 하에서도 95%이상의 인식률을 가진 것이 많다. 인식시스템의 경우, 성능 향상에 비례하여 응용분야도 복잡화 다양화되어가고 있다.

본 논문에서는 음소 단위의 에너지 파라미터를 이용하여 기준패턴을 그룹화하고 기존의 동적시간정합법(DTW)을 이용한 고립단어 인식시스템에 처리시간을 감소시켜 저가형 프로세서에서도 고속으로 동작할 수 있게 하도록 하였다. 저가형 프로세서를 사용한 음성인식이 가능하다면, 제안된 음성인식 알고리즘을 사용한 제품의 전체 가격 절감 효과를 기대할 수 있을 것이다.

Ⅱ. 특징 매개변수 검출 및 패턴정합

2.1 동적패턴정합(DTW)

패턴 정합 방법인 동적시간정합법(DTW: Dynamic Time-Warping)은 길이가 서로 다른 두 개의 자료에서 최적 정합 경로를 서로 비교할 수 있는 방법으로, 비교적 간단한 알고리즘과 최소의 하드웨어가 요구되므로 간단한 응용분야에 효율적으로 이용할 수 있다. 이 기술은 고립단어 인식에서 기원되었으나 연속음성 인식에 역시 적용할 수 있다. 그러나 동적 프로그래밍(DP: Dynamic Programing)으로 인해 계산량이 많고, 수많은 음성 내 변위를 수용할 수 있는 기준패턴 작성이 어려워 사용어휘가 제한되는 단점이 있다. 단어 음성의 시계열 패턴은 같은 화자가 발성하는 경우에도 발성에 따라서 지속시간이 변한다. 또한 패턴이 시간적으로 불균등하게 신축된다. 발성 속도가 변화해도 자음 부분, 자음으로부터 모음으로의 과도적 부분은 어느 정도 고유 길이를 가지고 있어서 비교적 변화가 적다. 그러나 모음부분은 큰 폭으로 신축이 일어난다. 그래서 시계열 패턴의 전 구간에서 비선형적인 시간 신축이 발생한다. 이러한 시간 신축현상을 반영하기 위한 방법이 동적시간신축방법(DTW)이다 (그림 1)[1].



(a) $P_1(n)$: 기준 패턴 LPC
(b) $P_2(m)$: 입력 패턴 LPC

그림 1. 동적시간정합

입력 단어 실험 패턴 T 는 단어 음성의 기준 패턴 R 을 특징 벡터의 시계열로 각각 식 (1)과 식 (2)로 표시

된다.

$$T = (a^1, a^2, a^3, \ldots, a^i, \ldots, a^I) \quad \cdots\cdots (1)$$

$$R = (b^1, b^2, b^3, \ldots, b^i, \ldots b^J) \quad \cdots\cdots (2)$$

특징 벡터 a^i 와 b^j의 거리 $d(a^i, b^j)$ 는 식 (3)과 같은 유클리드 거리로서 주어질 수 있다.

$$d(a^i, b^j) = \| a^i - b^j \|_2 = \left\{ \sum_{m=1}^{M} (a^i_m - b^j_m)^2 \right\}^{\frac{1}{2}} \quad \cdots\cdots (3)$$

거리를 생각하는 2개의 벡터의 시계열 a^i 와 b^j 를 대응시키는 것은 i 축과 j 축 상에서 평면상의 쌍을 식 (4)로 표시하고, 여기서 $F(k)$ 를 시간 변환 함수라고 부른다[2].

$$F(k) = f(1) f(2) \ldots f(k) \ldots f(K)$$
$$f(k) = f(i(k), j(k)), \; k = 1, 2 \ldots, k \quad \cdots\cdots (4)$$

2개의 특징 벡터의 시계열 T, R을 시간 변환 함수 $F(k)$에 의하여 대응 될 때의 T, R사이의 거리 $\overline{D}(T, R)$ 은 식 (5)와 같다.

$$\overline{D}(T, R) = \frac{\sum_{k=1}^{K} w(k) d(a^{i(k)}, b^{j(k)})}{\sum_{k=1}^{K} w(k)} \quad \cdots\cdots (5)$$

2개의 음성 패턴 T, R의 거리 $D(T.R)$은 시간 변환 함수 $F(k)$를 변화시킬 때의 $\overline{D}(T, R)$의 최소치로서 식 (6)으로 정의된다.

$$D(T.R) = \min_{F(k)} \overline{D}(T, R) = \frac{1}{N} \min_{F(k)} \sum_{k=1}^{K} w(k) d(a^{i(k)}, b^{j(k)}) \quad \cdots\cdots (6)$$

입력 단어 음성 패턴 T에 대해서, 모든 단어 음성의 기준 패턴 R과 DP 매칭을 행한 다음, 거리 $D(T, R)$을 최소로 하는 R의 패턴을 T의 패턴으로 선택한다. 즉 단어 음성 패턴 R에 의해서 표시되는 단어에 대응하는 것으로 판단한다.

2.2 음성구간 검출

음성 검출(end-point detection)은 음성을 발성한 주위 환경에 큰 영향을 받는다. 가장 이상적인 환경은 방음실과 같은 잡음이 존재하지 않는 밀폐된 공간이라고 할 수 있다. 그러나 이상적인 환경에서 항상 음성인식을 수행할 수는 없는 것이다. 그래서 어느 정도의 잡음이 존재하는 사무실 환경을 고려한다.

음성구간 검출의 정확성에 따라 인식 정확도에 큰 영향을 미치기 때문에 정확한 끝점검출(end-point detection)이 필요하게 된다. 또한 실시간 시스템에 사용하기 위해서는 전체 계산량을 크게 증가시키나 않는 효율적인 방법이어야 한다[3].

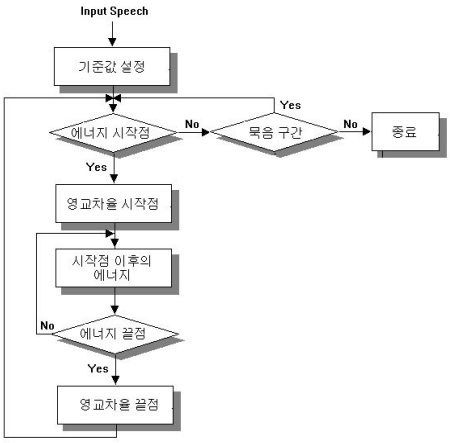


그림 2. 음성끝점 검출의 흐름도

간단하게 단구간 에너지(short-time energy)를 이용하여 에너지 값이 큰 부분은 음성구간으로, 작은 부분은 묵음구간(silence)으로 결정하는 방법이 있다. 이에 음성의 영교차율이 적은 부분을 검출하여 음성 구간을 검출하는 방법은 (그림 2)와 같다.

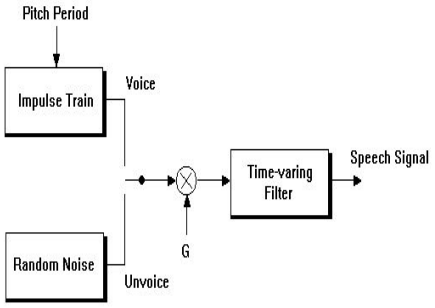


그림 3. 일반적인 음성 생성 모델

3.2 LPC(Linear Prediction Coding) 검출

음성의 생성모델은 그림 3과 같다. 우선 음원 성분인 유성음과 무성음을 생성한 뒤 여기신호 크기를 조절한 후 성도성분을 나타내는 시변환 필터를 거쳐 음성신호가 생성되는데 이때 성도를 나타내는 필터는 시변환 특성을 가지고 전극 구조이므로 그림 3과 같은 구조를 갖게 된다.

LPC 는 음성 신호처리에서 가장 널리 쓰이는 알고리즘 중 하나로 음성을 식 (7)과 같은 전극(all pole)모델로 가정하고 그에 따른 필터 계수를 이용하여 음성 신호를 모델링 한다.

이 LPC 계수로 구성되는 필터는 전극특성으로 가정하여 음성이 어떻게 생성되는가를 분석하여, 성도의 특성을 모델링하게 된다. 또한 실제 구현 시 적용이 쉽기 때문에 많이 사용되는 알고리즘이다[1][4].

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \quad \cdots\cdots (7)$$

III. 제안된 음성인식 알고리즘

현재의 음성인식시스템은 응용분야에 따라 동적패턴정합(DTW), 벡터양자화(VQ), 은닉마코프 모델(HMM), 신경망(NN) 등의 다양한 방법으로 개발되어있다[5][6][7]. 그 중에서 고립단어인식 분야에서는 알고리즘 복잡성과 하드웨어로 구현이 비교적 간단한 동적패턴정합 방법이 많이 이용된다. 하지만 패턴동적정합은 정합방식을 모든 기준 단어에 대해 반복 수행함으로 알고리즘 복잡성에 비해 처리시간이 긴 단점이 있다[5][6].

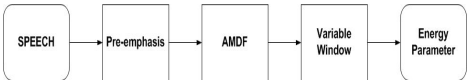
본 논문에서 제안한 고립단어인식 알고리즘은 다음과 같다.

기존 DTW를 사용한 고립단어인식시스템의 계산량 감소를 위해 기준패턴 음성의 최초 음절을 에너지에 따라 유무성음을 분리하고, 구분된 유성음 구간을 라벨링하고, 라벨링된 값에 따라 기준패턴을 네 개의 그룹으로 분리한다. 단어인식 수행단계에서 입력된 테스트패턴 음성의 최초 음절 에너지를 추출하여 그 레벨에 맞는 그룹에서 DTW를 수행하여 DTW 수행 처리시간을 감소시켰다.

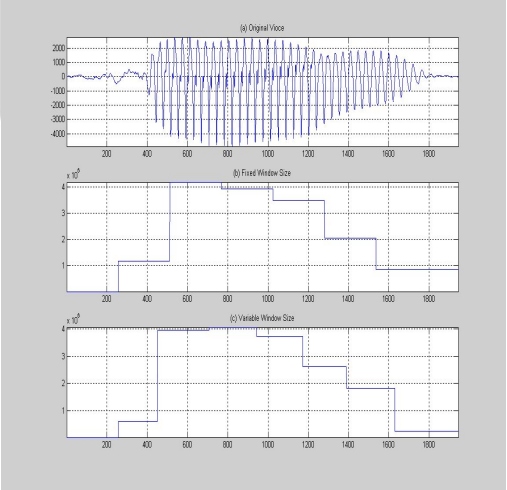
본 논문의 음성 라벨링 단계는 음성의 구간 검출 및 에너지 파라미터 추출 단계에서 정확한 파라미터 검출을 전제로 하므로 이를 보안하기 위해 피치 주기에 따른 가변윈도우를 사용하였다. 기존의 방법은 윈도우 사이즈 256프레임으로 고정시켜 에너지를 구함으로써 정확한 프레임 에너지를 구할 수 없었다. 따라서 피치주기를 먼저 구하고 그 주기에 200 프레임에서 300프레임 사이에서 윈도우 크기를 결정함으로써 윈도우 영향이 제거된 에너지를 구하는 방법을 제안하였다. 제안된 가변 윈도우를 사용하여 에너지를 구하는 방법은 식 (8)과 같이 표현된다.

$$E_n = \frac{\sum_{i=0}^{\Pi tch\_peroid} E(i)}{\Pi tch\ period} \quad \cdots\cdots (8)$$

제안한 가변 윈도우 흐름도는 (그림 4)와 같다.



AMDF : Average Magnitude Difference Function
그림 4. 가변윈도우를 사용한 에너지 추출 흐름도



(a) 입력 음성 (b) 고정 윈도우를 사용한 단구간 에너지
(c) 가변 윈도우를 사용한 단구간 에너지

그림 5. 가변윈도우를 사용한 에너지추출 결과파형

(그림 5)는 200 프레임에서 300프레임 사이에서 피치주기에 따라 가변적으로 결정된 윈도우를 사용한 결과 파형

이다. (그림 5)의(a)는 입력된 음성(original voice)이고, (그림 5)의 (b)는 256프레임의 고정윈도우를 사용한 결과이고, (그림 5)의 (c)는 가변윈도우를 사용한 결과이다. (그림 5)의 (c)가 그림 5의 (b)보다 정확한 에너지 파라미터와 무성음구간을 검출해내는 것을 알 수 있다.

(그림 6)은 본 논문에서 제안한 에너지 라벨링을 이용한 음성인식 시스템의 전체적인 구조이다.

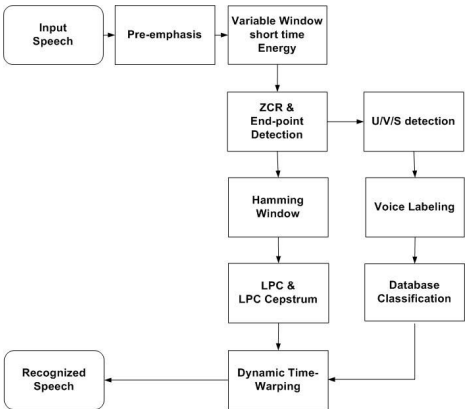


그림 6. 제안된 음성인식 시스템

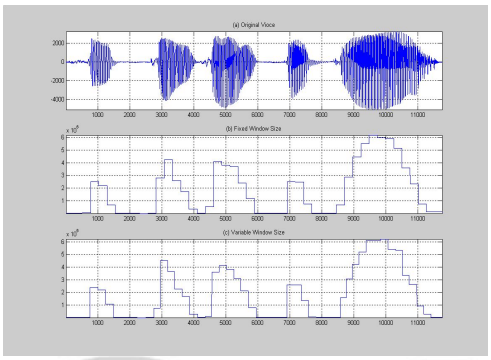
IV. 실험결과

제안한 알고리즘의 모의실험을 위해 PC에 마이크가 장치된 16비트 A/D변환기를 장착시켰다. 알고리즘을 구현 도구로 Matlab 6.0을 사용하여 파형분석을 하였고, 전체 인식률과 속도 측정을 위해Win32api를 사용하여 인식기를 구현하였다. 입력 음성 시료는 8khz로 샘플링하고 16bit로 양자화하여 사용하였다. 단어인식에 사용되어진 특징벡터는 10차 LPC(Linear Prediction Coding) 계수를 사용하였고, 특징벡터 인식을 위해 DTW(Dynamic Time Warping)를 사용하였다.

4.1 윈도우 영향이 제거된 에너지 매개변수 추출

기존의 단시간 에너지 파라미터 추출(short-time energy parameter extraction) 방법은 윈도우 사이즈를 고정시켜 에너지 파라미터를 구함으로 음성 전이구간에서 정확한 프레임 에너지를 구할 수 없다. 따라서 본 논문에서는 피치주기를 먼저 구한 후 그 주기에 따라 에너지를 구하는 방법을 제안하였다.

(그림 7)은 가변윈도우(variable windows size)를 사용한 모의실험 결과이다. (그림 7)의(a)는 입력된 음성(original voice)이고, (그림 7)의(b)는 256프레임의 고정된 윈도우를 사용한 결과이며, (그림 7)의(c)은 피치 주기에 따른 가변 윈도우를 사용한 결과이다. 제안된 가변윈도우를 사용한 방법은 입력 음성의 전이구간에서 정확한 에너지 파라미터를 구하는 결과를 보인다.

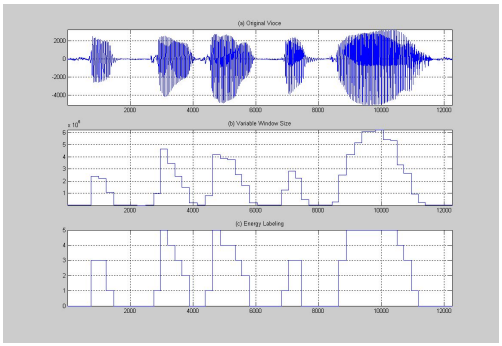


(a) 입력 음성 (b) 고정 윈도우를 사용한 단구간 에너지
(c) 가변 윈도우를 사용한 단구간 에너지

그림 7. 제안된 에너지 추출 방법의 결과 파형

4.2 검출된 에너지를 이용한 기준패턴 유·무성음 분류 및 기준패턴 그룹화

제안한 가변윈도우를 사용한 에너지 파라미터를 이용하여 음성의 U/V/S를 구별하고 구별된 유성음 구간에서 에너지를 라벨링하여 기준패턴을 그룹화 하였다.



(a) 입력 음성 (b) 검출된 에너지 파라미터
(c) 에너지 라벨링

그림 8. 입력된 음성의 에너지 라벨링

(그림 8)은 유/무성음 분리와 유성음 구간에서의 에너지 라벨링 결과이다. 실험에 사용되어진 음성은 8㎑로 샘플링 되고 16bit로 양자화된 raw화일이다.

4.3 검색된 그룹 내에서의 음성인식 결과

실제 인식 수행 시에는 라벨링된 에너지에 따라 미리 입력된 기준패턴을 4개로 그룹화하여 입력된 음성의 에너지 크기에 따라 미리 라벨링된 패턴 그룹 내에서 DTW를 수행했다.

실험은 밀폐된 일반 실험실 환경에서 10명의 화자가 각각 20개의 지정된 단어를 발성한 후 다시 테스트 음성을 발성하였다. 이 과정을 지속적으로 수행하여 인식률과 처리속도를 계산하였다.

실험결과 인식률에서는 기준음성의 화자와 실험음성의 화자가 동일한 경우에 기존의 방법과 제안된 방법이 약 92% 정도로 비슷한 성능을 나타내었고, 기준음성의 화자와 실험음성의 화자가 다른 경우에는 기존의 방법이 약 85% 정도이고 제안된 방법이 81%정도로 제안한 방법이 4%정도의 성능 저하를 나타내었다. 하지만 처리시간의 면에서는 동일화자와 복수화자의 경우 모두 제안된 방법에서 76%정도로 약 25%정도의 연산량을 감소시켰다.

입력된 음성의 첫 음절의 에너지 값이 1.5×10^{6}미만인 경우는 무성음을 나타내고 에너지 값이 1.5×10^{6}이상인 경우는 유성음을 나타낸다. 또 에너지 값이 1.5×10^{5}미만인 경우는 무음(silence)구간으로 나타났다. 구해진 유성음 구간을 3.0×10^{6}, 4.0×10^{6}, 5.0×10^{6}으로 라벨링 하였다.

V. 결 론

현재 음성인식시스템은 응용분야에 따라 동적패턴정합(DTW), 벡터양자화(VQ), 은닉마코프 모델(HMM), 신경망(NN) 등의 다양한 방법으로 개발되었다. 그중에서 고립단어인식 분야에서는 알고리즘의 복잡성과 하드웨어로 구현이 비교적 간단한 동적패턴정합 방법이 많이 이용되고 있다. 하지만 패턴동적정합은 반복적인 정합방식에 의존함으로 알고리즘 복잡성에 비해 처리시간이 길다는 단점이 있다.

본 논문에서는 음소단위로 추출된 에너지 파라미터를 이용하여 에너지를 라벨링하고 라벨링 된 값에 따라 입력음성을 그룹화 하였다. 그리고 동적패턴정합 수행 시 입력된 실험음성에서 검출된 에너지의 크기에 따라 선택되어진 라벨의 그룹 내에서 DTW를 수행시켜 처리시간을 단축시켰다. 결과적으로 기존의 DTW를 사용한 고립단어 인식시스템에 연산량을 감소시켜 저가형 프로세서에서도 고속으로 동작할 수 있게 하고자 하였다.

저가형 프로세서를 사용하여 음성인식을 수행할 수 있다면, 이 음성인식 알고리즘을 사용한 제품에 전체 가격을 절감시키는 효과를 기대할 수 있어 더 넓은 분야에서 음성인식기술을 접해볼 수 있을 것이다.

참고문헌

[1] S. Furui, “Digital Speech Processing, Synthesis and Recognition,” Marcel Dedder, Inc., 1992.

[2] L. R. Rabiner & R.W.Schafer, “Digital Processing of Speech Signa,l” Prentice-Hall, Englewood Cliffs, N.J., U.S.A, 1978.

[3] L. R. Rabiner & Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall AT&T, U.S.A, 1993.

[4] 이기희, “선형 변환망을 이용한 화자적응 음성인식”, 한국컴퓨터정보학회 논문지, 제5권, 제2호, pp.90-97, 2000.

[5] 조태수, “윈도우 영향이 제거된 에너지 파라미터에 관한 연구,” 대한전자공학회 하계종합학술대회, 2001.

[6] 우상욱, 권승호, 한수영, 이동규, 이두수, “어휘 그룹화를 이용한 음성인식시스템의 성능향상에 관한 연구”, 대한전자공학회 하계종합학술대회, 2003.

[7] 지진구, 윤성일, “음성을 이용한 화자 검증기 설계 및 구현”, 한국컴퓨터정보학회 논문지, 제5권, 제3호, pp.91-98, 2000.

저 자 소 개

한 수 영

2004년 2월 한양대학교 전자공학과 공학박사

현재 안양대학교 컴퓨터학과 전임강사

김 홍 렬

1997년 8월 한양대학교 전자공학과 공학박사

현재 동서울대학 컴퓨터정보과 조교수

이 기 희

1996년 2월 한양대학교 전자공학과 공학박사

현재 동서울대학 컴퓨터정보과 부교수